SUPPORTING INFORMATION

Discovery and visualization of uncharacterized drug-protein adducts using mass spectrometry.

Michael Riffle[1†], Michael R. Hoopmann[2†], Daniel Jaschob[1], Guo Zhong[3], Robert L. Moritz[2], Michael J. MacCoss[4], Trisha N. Davis[1], Nina Isoherranen[3], Alex Zelter[1*]

Departments of [1]Biochemistry, [3]Pharmaceutics and [4]Genome Sciences, University of Washington, Seattle, WA; [2]Institute for Systems Biology, Seattle, WA. [†]M.R. and M.H. contributed equally to this work. *Corresponding author: Alex Zelter, University of Washington, Department of Biochemistry, 1959 NE Pacific Street, HSB, Box 357350, Seattle, Washington, 98195, USA. Email: azelter@uw.edu

Table of Contents

SUPPORTING INFORMATION 1

Supplementary Methods 5

Reagents 5

Drug Incubations 5

Mass Spectrometry 5

Sample Preparation *5*

Chromatography *6*

MS Instruments and Data Acquisition Settings *6*

MS Data Processing and Database Searching *7*

Supplementary Note 1: Magnum 11

Overview 11

Reading and Processing Inputs 11

Table S1: Select parameters relevant to open modification searches in Magnum *13*

Open Modification Database Searching 14

Results Scoring 15

MS-labile versus stable adducts 16

Restriction of open modification mass to specific residues in Magnum 16

Adduct reporter ions 16

Error estimation and PSM utilities in Magnum 16

Figure S2: Distribution of –log base 10 e-values of targets and decoys from a single Magnum example search. *17*

Figure S3: Histogram of –log base 10 e-values of decoys from Magnum from 24 example searches. *18*

Magnum search engine speed and search space recommendations....18
Supplementary Note 2: Gold standard dataset for evaluation of xenobiotic-protein adduct discovery....19
Table S2: Raw files used for creation of gold standard data. 19
Figure S4: The structure of (a) dicloxacillin and (b) flucloxacillin after forming adducts on a lysine primary amine. 20
Table S3: Number of MS/MS spectra that contain β-lactam antibiotic reporter ions at the defined m/z and intensities. 21
Figure S5: Workflow for the creation of 3 gold standard datasets. 22
Table S4: Number of MS/MS spectra that contain β-lactam antibiotic reporter ions at the defined m/z and lower intensities than searched for in Table S3. 23
Table S5: Number of MS/MS spectra that contain a single defined 469.0266 (dicloxacillin) or 453.0561 (flucloxacillin) adduct at a Percolator q ≤ 0.05 based on a comet closed search. 23
Figure S6: Histograms showing the distribution of open modification masses reported 24
Figure S7: 25
Figure S8: Precision recall plot of adduct masses reported at 1% FDR by 7 open search algorithms for MS/MS spectra definitively determined to result from a peptide containing a single +469 Da (dicloxacillin) or +453 Da (flucloxacillin) modification. 26
Figure S9: Histograms showing the distribution of open modification masses reported by each algorithm for the 763 gold standard spectra generated from dicloxacillin treated samples using method 1....27
Figure S10: Histograms showing open modification masses reported by each algorithm for the 1,248 gold standard spectra generated from dicloxacillin treated samples by method 2....27
Figure S11: Histograms showing the distribution of open modification masses reported by each algorithm for the 1,702 gold standard spectra generated from flucloxacillin treated samples using method 2....28
Table S6: Gold standard results returned by open-pFind at 1% FDR. 29
Additional Benchmarking of Magnum and MSFragger using Phosphopeptides 29
Figure S12: Comparison of phospho-modifications identified by closed comet searching (defined phospho-modification allowed on S, T or Y) versus Magnum and MSFragger open searches (60-500 Da allowed on any amino acid). 30
Supplementary Note 3: Limelight 31
Description 31
Experiment Builder 31
Figure S13: Iteratively building an experiment using the experiment builder. 32
Score filters and cutoffs 32
Single protein view....32

Peptide view 33
Protein view 33
Modification view 33
Visualizations and Transformations *34*
Two-tailed test of proportions *34*
Supplementary Note 4: Development and validation of adduct discovery pipeline using dicloxacillin, flucloxacillin and HSA 36
Open modifications unrelated to exposure 36
Figure S14: Open modification masses in untreated, purified human serum albumin as identified by 7 different open search algorithms and visualized by Limelight. *36*
Figure S15: The number of PSMs at 1% FDR resulting from searching 6 untreated and dicloxacillin and flucloxacillin exposed HSA datasets with open and closed search algorithms. *37*
Discovery of dicloxacillin/flucloxacillin adducts in HSA 37
Figure S16: Open modification masses identified by Magnum or MSFragger in untreated, dicloxacillin treated or flucloxacillin treated human serum albumin (HSA). *38*
Figure S17: A two-tailed test of proportions performed within Limelight identifies treatment specific adducts in HSA. *39*
Table S7: Dicloxacillin and flucloxacillin adducted peptides identified by Magnum In purified HSA. .*40*
Figure S18: *43*
Figure S19: Annotated MS and MS/MS spectrum of flucloxacillin adducted HSA peptide identified by Magnum. *44*
Discovery of dicloxacillin adducts in human plasma 45
Figure S20: Open modification analysis of untreated and dicloxacillin treated human plasma. PSMs were identified by Magnum and visualized by Limelight. *45*
Table S8: Dicloxacillin adducted peptides identified by Magnum in human plasma. *46*
Supplementary Note 5: Localization of dicloxacillin and flucloxacillin adducts 47
Table S9: Dicloxacillin and flucloxacillin adduct localization determined by Magnum or MSFragger plus PTMProphet in 2,979 gold standard MS/MS spectra. *47*
Supplementary Note 6: CYP3A4/Raloxifene analysis 48
Figure S21: Bioactivation of raloxifene by CYP3A4 and representative adducts formed with cysteine, tryptophan or tyrosine. *48*
Initial Experiments and Skyline Quantification 48
Table S10: A two-tailed test of proportions identifies a 471 Da raloxifene specific adduct mass in CYP3A4 and P450-reductase. *49*
Figure S22: Identification of novel raloxifene adducts in CYP3A4 and P450-reductase *50*

Table S11: All locations identified as modified by 471 Da adduct masses by Magnum-CWY in CYP3A4 and P450-reductase in initial experiments by ≥ 2 PSMs. 51

Figure S23: Total normalized ion signal in treated and untreated samples, quantified in Skyline 52

Table S12: Full peptide sequences corresponding to the peptide letter abbreviations used in Figure S23. 53

Figure S24: Total PSMs identified in CYP3A4 and P450-reductase at each location found by Magnum-CWY searches as modified by a 471 Da adduct mass in initial experiments by ≥ 2 PSMs...54

Figure S25: Total normalized ion signal quantified in Skyline for all locations identified in CYP3A4 and P450-reductase at each location found by Magnum-CWY searches as modified by a 471 Da adduct mass in initial experiments by ≥ 2 PSMs. 55

Several raloxifene adducts result in multiple distinct chromatographic peaks....56

Figure S26: Precursor ions resulting corresponding to the W126 peptide (SAISIAEDEEW[471]KR) elute as 4 distinct chromatographic peaks. 56

Figure S27: Precursor ions resulting corresponding to the C98 peptide (EC[471]YSVFTNR) elute as 3 distinct chromatographic peaks. 57

Figure S28: Precursor ions resulting corresponding to the C468 peptide (VLQNFSFKPC[471]K) elute as 2 distinct chromatographic peaks....57

Figure S29: Precursor ions resulting corresponding to the C58 peptide (GFC[471]M[16]FDMEC[57]HKK) elute as 1 single distinct chromatographic peak.....58

Further raloxifene experiments 58

Table S13: Magnum identifies multiple 471 Da protein adducts in CYP3A4 and P450-reductase after exposure to raloxifene. 59

Figure S30: Identification of novel raloxifene adducts in (a) CYP3A4 and (b) P450-reductase. 60

Figure S31: Distribution of Magnum identified open-mass modifications within the 471 Da bin in all untreated and raloxifene treated samples combined.....61

Verification of raloxifene adducts by standard closed searching 61

Table S14: Comet verification of Magnum-identified 471 Da adducts presented in Table S12. 62

Key protein sequences: 63

Table S15: Protein sequences of human serum albumin (HSA) plus the heterologously expressed proteins CYP3A4 and rat P450 reductase proteins. 63

Supplementary References: 64

Supplementary Methods

Reagents

Human serum albumin (HSA), flucloxacillin sodium, dicloxacillin sodium, and raloxifene were purchased from MilliporeSigma (Burlington, MA). Sequencing Grade Modified Trypsin was purchased from Promega Corporation (Madison, WI), product number V5111. Rat P450 reductase was expressed and purified as described previously[1]. Purified recombinant human CYP3A4 and the liposome stock containing l-α-Dilauroyl-sn-glycero-3-phosphocholine (DLPC), l-α -diloleoyl-sn-glycero-3-phosphocholine (DOPC), and l- α -dilauroyl-sn-glycero-3-phosphoserine (DLPS) (1:1:1, w/w/w per mL) were gifts from Dr. William Atkins, University of Washington. The expression and purification of human recombinant CYP3A4 was performed as described previously[2].

Drug Incubations

Human serum albumin (0.5 mg/mL) was incubated with and without 100 μM antibiotics (flucloxacillin and dicloxacillin) in 100 mM potassium phosphate buffer (pH 7.4) at 37°C for 16 hours. The total volume was 200 μL. The control group contained no antibiotics. After incubation unreacted antibiotics were removed by buffer exchange into 100 mM potassium phosphate buffer (pH7.4) using protein desalting spin columns (ThermoFisher part number 89849) according to manufacturer's instructions. Incubations were performed in duplicates.

Recombinant human CYP3A4 (3 μM) was incubated with 6 μM rat P450-reductase, 2 mM NADPH in the absence and presence of 200 μM raloxifene in buffer containing 20 μg/mL liposomes [DLPC, DOPC, DLPS (1:1:1, w/w/w per mL)], 0.1 mg/mL CHAPS, 3 mM glutathione, 30 $MgCl_2$ and 50 mM potassium HEPES buffer (pH 7.4). The control group contained no raloxifene. Incubations were done in singlet at 37°C for 1 hour in a total volume of 200 μL and the reaction was started by adding NADPH.

For plasma experiments fresh blood was collected and left on ice for 1 hour and was then centrifuged at 1000 g for 10 minutes at 4°C. Plasma was collected and a portion of undiluted plasma (60 μL) was flash-frozen immediately using liquid nitrogen. The rest of undiluted plasma (60 μL in each sample) was spiked with dicloxacillin (2 μL of 50 mM stock in water) and incubated at 37°C for 4 hours. Samples were flash-frozen using liquid nitrogen after incubation and were stored at -80°C for later MS analysis.

Mass Spectrometry

Sample Preparation

"Extra digest" raloxifene and CYP3A4: Aliquots (100 μL) of control or drug treated CYP3A4 incubation reaction mixture (63 μg total protein) were incubated at 75°C for 30 minutes. Samples were brought to 5.5 mM TCEP and reduced for 1 hour at 60°C, cooled to room temperature, brought to 6 mM iodoacetamide and alkylated by incubating at room temperature in the dark for 30 min prior to a 16-hour trypsin digestion at 37°C.

All CYP3A4 samples were cleaned up after digestion by solid phase extraction using Oasis MCX cartridges (1 cc/30 mg cartridges, Waters corporation, product number 186000782) according to manufacturer's instructions. The resulting eluate was dried in a speed vac and reconstituted into 100 μL 0.1% trifluoroacetic acid, 2% acetonitrile in water before transfer to autosampler vials and storage at -80°C prior to MS analysis.

For human plasma samples, undiluted untreated or dicloxacillin treated plasma (2 µL) was diluted 1:10 by adding 18 µL of 50 mM ammonium bicarbonate in water. Diluted plasma (4.1 µL ~33 µg protein) was further diluted by adding 43.2 µL of 50 mM ammonium bicarbonate. Yeast enolase, 2.6 µL at 200 ng/µL (Sigma, Cat# E6126-500UN) was added as a "protein process control" (800 ng enolase per 50 µg sample protein). PPS Silent Surfactant (expedeon.com, Cat# 21011) was added (2.5 µL of 2% PPS in 50 mM ammonium bicarbonate) plus 1.4 µL of 200 mM TCEP. Samples were then incubated at 95°C for 5 minutes, cooled to 60°C and reduced for 1 hour in an Eppendorf Thermomixer with shaking (1000 rpm). Samples were alkylated for 30 minutes in the dark at room temperature by adding 1.3 µL of 250 mM iodoacetamide. After alkylation DTT was added (0.57 µL 500 mM DTT) prior to tryptic digestion was at a 1:15 (enzyme:substrate) for 6 hours at 37°C in an Eppendorf Thermomixer with shaking (1000 rpm). After digestion samples were acidified with 250 mM HCl (final concentration), incubated on the bench for 1 hour at room temperature and spun at max speed in a benchtop microfuge for 10 mins. Supernatant was transferred to autosampler vials and stored at -80°C until use.

Chromatography

Sample digest (2 µL ~1 µg) was loaded by autosampler onto a 150 µm Kasil fritted trap packed with 2 cm of ReprosilPur C18AQ (3 µm bead diameter, Dr. Maisch) at a flow rate of 2 µL per min. Desalting was performed with 8 µL of 0.1% formic acid plus 2% acetonitrile and the trap was subsequently brought online with a Self-Packed PicoFrit Column (New Objective part number PF360-75-10-N-5, 75 µm i.d.) packed with 30 cm of ReprosilPur C18AQ (3 µm bead diameter, Dr. Maisch) mounted in an in-house constructed microspray source and placed in line with a Waters Nanoacquity binary UPLC pump plus autosampler. Peptides were eluted from the column at 0.25 µL/min using an acetonitrile gradient. A standard gradient was used in all runs except those otherwise specified. A higher acetonitrile gradient was used for select CYP3A4 samples (labeled “highB” in results). Gradient details are described in Supplementary Methods.

The standard chromatography gradient used in all runs except those otherwise specified consisted of the following steps: (1) 0-20 mins; 2-7.5% B; flow 0.25 µL/min; (2) 20-100 mins; 7.5-25% B; flow 0.25 µL/min; (3) 100-140 mins; 25-60% B; flow 0.25 µL/min; (4) 140-145 mins; 60% B; flow 0.25 µL/min; (5) 145-146 mins; 60-95% B; flow 0.25 µL/min; (6) 146-151 mins; 95% B; flow 0.45 µL/min; (7) 151-152 mins; 95-2% B; flow 0.45 µL/min; (8) 152-179 mins; 2% B; flow 0.45 µL/min; (9) 179-180 mins; 2% B; flow 0.25 µL/min.

A higher acetonitrile gradient was used for select CYP3A4 samples (labeled “highB” in results) that consisted of the following steps: (1) 0-20 mins; 2-15% B; flow 0.25 µL/min; (2) 20-70 mins; 15-60% B; flow 0.25 µL/min; (3) 70-135 mins; 60-95% B; flow 0.25 µL/min; (4) 135-141 mins; 95% B; flow 0.5 µL/min; (5) 141-142 mins; 95-2% B; flow 0.5 µL/min; (6) 142-164 mins; 2% B; flow 0.5 µL/min; (7) 164-165 mins; 2% B; flow 0.25 µL/min. For all gradients, buffer A was: 0.1% formic acid in water and buffer B was 0.1% formic acid in acetonitrile.

MS Instruments and Data Acquisition Settings

All samples, except for human plasma, were run on a Thermo Fisher Scientific QExactive HF in data dependent mode using the following settings. A maximum of 20 tandem MS (MS/MS) spectra were acquired per MS spectrum (scan range of *m/z* 400–1,600). The resolution for MS and MS/MS was 60,000 and 15,000, respectively, at *m/z* 200. The automatic gain control targets for MS and MS/MS were set to set to a nominal value of 3e6 and 2e5, respectively, and the maximum fill times were 50 and 100 ms, respectively. MS/MS spectra were acquired using an isolation width

of 2.5 *m/z* and a normalized collision energy of 27. MS/MS acquisitions were prevented for +1, ≥+6 or undefined precursor charge states. Dynamic exclusion was set for 5 s. MS spectra were collected in profile mode and MS/MS spectra were centroided.

For human plasma samples a Thermo Fisher Scientific Exploris 480 was used with the following settings. DDA was run using a 3 second cycle time between each MS spectrum. Tandem MS (MS/MS) spectra were acquired with a scan range of *m/z* 400–1,600. The resolution for MS and MS/MS was 60,000 and 15,000, respectively, at *m/z* 200. The normalized automatic gain control targets for MS and MS/MS were set to 300% and 100%, respectively, and the maximum injection times were were set to auto. MS/MS spectra were acquired using an isolation width of 2.5 *m/z* and a normalized collision energy of 27. MS/MS acquisitions were prevented for +1, ≥+6 or undefined precursor charge states. Dynamic exclusion was set for 20 s. MS and MS/MS spectra were collected in centroid mode.

MS Data Processing and Database Searching

Acquired spectra were converted into mzML (for input to all algorithms except MODa) or mzXML (for input to MODa) using ProteoWizard's msConvert[3]. Proteins present in the samples were identified using Comet[4] by standard closed searching against the entire human proteome, for HSA and plasma samples, or the E. coli proteome, for CYP3A4 samples, plus common contaminants (https://www.thegpm.org/crap/). For CYP3A4 samples protein sequences for the heterologously expressed proteins CYP3A4 and P450-reductase (Table S14) were appended to the search database. A q-value was assigned to each PSM through analysis of the target and decoy distributions using Percolator[5]. Smaller databases were made for subsequent open searching consisting only of proteins identified in initial comet searches by at least 3 peptides with a Percolator assigned q-value of ≤ 0.01. Decoy databases consisted of the corresponding set of reversed protein sequences and were provided to algorithms requiring pre-generated decoy sequences. All data reported are filtered at a false discovery rate (FDR) of 1% unless otherwise stated. Detailed descriptions of all database search procedures for each algorithm can be found in Supplementary Methods.

Comet (closed). Searches were performed on untreated, flucloxacillin and dicloxacillin treated HSA samples for generation of gold standard results (Supplementary Note 2) and as positive controls in HSA dicloxacillin and flucloxacillin gold standard searches by defining the known dicloxacillin or flucloxacillin adduct masses as variable modifications allowed on lysine (469.0266 or 453.0561, respectively). Comet[4] version 2019.01 rev. 0 was used for these searches configured with a 15 ppm precursor mass tolerance, a fragment bin tolerance of 0.03, and an isotope error of 3. Percolator[5] version 3.02.1 was used to assign q-values to comet generated PSMs.

Comet (500 Da wide precursor). Searches were performed as for closed searches above, but with a 500 amu precursor mass tolerance and an isotope error of 1. Percolator version 3.02.1 was used to assign q-values to comet generated PSMs.

Comet-PTM (open). Searches were done using Comet version "PTM 2016.01 rev. 2". Precursor mass tolerance and delta_outer_tolerance were set to 500 amu, delta_inner_tolerance was 0.8 and fragment_bin_tol was 0.02. FDR was calculated by the Limelight XML converter as described below.

Magnum (open). Searches were performed using Magnum version 1.0-dev.11 for all searches except human plasma and the phospho-peptide benchmarking described in Supplementary Note 2. The latter searches used Magnum version 1.0.0-alpha5. Open modifications between 60 and 500 Da were allowed on all amino acids (adduct_sites = ARNDCQEGHILKMFPOSUTWYV), K only (adduct_sites = K), C only (adduct_sites = C), CWY only (adduct_sites = CWY). Defined variable modifications were oxidation of methionine and carbamidomethylation of cysteine. E value depth was set to 10000. Precursor mass tolerance was 15 ppm and isotope error was set to 1. For dicloxacillin and flucloxacillin HSA searches reporter ions were defined as follows, using a reporter ion threshold of 5: 160.04; 311.00; 470.03; 295.03; 454.06. For plasma searches reporter ions were: 160.04; 311.00; 470.03. Percolator version 3.05.0 was used to assign q-values to Magnum generated PSMs.

MetaMorpheus (open). Searches were performed using MetaMorpheus version 0.0.308. An initial calibrate task was performed using a precursor mass tolerance of 15 ppm and a product mass tolerance of 25 ppm. Post calibration searches were performed using precursor and product mass tolerances of 5 and 20 ppm, respectively. The modern search option was specified, and open-mass differences of -187 and up were allowed. Native q-values calculated by MetaMorpheus were used for FDR filtering.

MODa (open). Searches were done using version 1.62 allowing open modifications of 60-500 Da with a precursor mass tolerance of 15 ppm and a product mass tolerance of 0.03 Da. Variable modifications cannot be defined in MODa. A fixed modification for carbamidomethylation of cysteine was used during analysis of dicloxacillin and flucloxacillin HSA samples as this resulted in increased sensitivity in gold standard searches. For raloxifene searches the native mass of cysteine was used (no fixed carbamidomethylation of cysteine) as raloxifene is known to modify cysteine; instead, open modifications were allowed between 55-500 Da to incorporate the possibility of carbamidomethylation of cysteine. For all searches, data were presented for searches where blind mode was set to 1, allowing 1 open modification per peptide. This was found to be much more sensitive than blind mode 2, which allows an arbitrary number of modifications per peptide. Instrument was set to ESI-TRAP and high resolution was set to on. FDR was calculated by the Limelight XML converter as described below.

MSFragger (open). Searches were performed using version 2.3 allowing open modifications between 60 and 500 Da. Precursor and fragment mass tolerance was set to 15 ppm and calibrate mass was set to 2. Isotope error was 0/1 and localize delta mass was 1. MSFragger output was processed with PeptideProphet[6] and PTMProphet[7] (TPP v5.2.1-dev Flammagenitus, Build 202003241419-8041) for error rate estimation and open modification localization using the following options for PeptideProphetParser: ACCMASS DECOYPROBS DECOY=random NONPARAM MASSWIDTH=520 MINPROB=0. PTMProphetParser was run using MASSDIFFMODE MINPROB=0. FDR was calculated by the Limelight XML converter as described below.

open-pFind (multinotch). Searches were done using version EVA.3.0.11 with a precursor mass tolerance of 15 ppm and a fragment tolerance of 0.03 Da. Mixture spectra was true, precursor score model was normal and a threshold of -0.5 was used. Peptide mass was between 600 and 6000 and length between 6 and 100. Open search was set to true. FDR was calculated by the Limelight XML converter as described below.

All searches were performed requiring fully tryptic peptides allowing either 2 or 3 missed cleavages and defining variable modifications of oxidation (15.9949) of methionine and carbamidomethylation (57.021464) of cysteine except MODa which does not allow for defined variable modifications. Search databases consisted of all proteins detectable in the sample by searching spectra against whole-proteome databases using comet. PSMs were processed with Percolator and proteins identified by ≥ 3 PSMs with a Percolator assigned q-value ≤ 0.01 were included in a smaller protein database used for open modification searches. Algorithms that required pre-generated decoys sequences were given the reversed sequence of each target sequence. For each search performed, software version numbers and complete configuration files are available for download via Limelight or via ProteomeXchange (see below).

False Discovery Rate (FDR) calculation. When a software pipeline produces an FDR or q-value associated with a PSM, that value was used to filter the data at a 0.01 FDR or q-value threshold directly. The software pipelines below do not produce a PSM-level score analogous to a FDR or q-value. This value was therefore calculated and associated with every PSM as follows.

MSFragger (post processed by TPP). This pipeline produces a probability score associated with each PSM that represents the probability that the identified peptide is correctly associated with the given spectrum. So, to calculate the predicted FDR an estimation for the number of incorrectly identified PSMs can be calculated as the sum of 1 minus this probability for all PSMs with a given probability or better. This sum can be divided by the total number of PSMs with a given probability or better to obtain an estimate of the FDR. The code may be viewed on GitHub at https://github.com/yeastrc/limelight-import-msfragger-tpp.

Comet-PTM. All unique E-value scores (reported by Comet-PTM) are sorted from best to worst. This list is then iterated over and for each score a sum is calculated for the total number of target hits and the total number of decoy hits with that score or better. An FDR is calculated for the given score as the total number of decoy hits divided by the total number of decoy hits plus total number of target hits. Then the FDR for any previously processed score (i.e., better scores) is changed to the minimum of its existing FDR or the current score's FDR. The code may be viewed on GitHub at https://github.com/yeastrc/limelight-import-cometptm.

MODa. All unique probability scores (reported by MODa) are sorted from best to worst. This list is then iterated over and for each score a sum is calculated for the total number of target hits and the total number of decoy hits with that score or better. An FDR is calculated for the given score as the total number of decoy hits divided by the total number of decoy hits plus total number of target hits. Then the FDR for any previously processed score (i.e., better scores) is changed to the minimum of its existing FDR or the current score's FDR. The code may be viewed on GitHub at https://github.com/yeastrc/limelight-import-moda.

Open-pFind. All unique final scores (reported by open-pFind) are sorted from best to worst. This list is then iterated over and for each score a sum is calculated for the total number of target hits and the total number of decoy hits with that score or better. An FDR is calculated for the given score as the total number of decoy hits divided by the total number of decoy hits plus total number of target hits. Then the FDR for any previously processed score (i.e., better scores) is changed to the minimum of its existing FDR or the current score's FDR. The code may be viewed on GitHub at https://github.com/yeastrc/limelight-import-open-pfind

Quantification of CYP3A4 and P450-reductase adducts. Peptides were quantified by integrating and summing the area under the curve of the peptide elution for specific transitions

using Skyline[8,9]. All MS1 transitions corresponding to residues identified in initial Magnum-CWY searches as having a 471 Da modification in ≥ 2 PSMs (Table S11) were quantified as previously described[9]. Additionally, 5 un-modified CYP3A4 and 5 unmodified P450-reductase peptides were chosen for normalization of the data. These contained no raloxifene modifiable residues (e.g. C, W, Y) and no tryptic ragged ends (e.g. C-terminal KK, RR, KR or RK). The total extracted MS1 transition area of each normalization peptides within each experiment was summed. For each 471 Da modified transition the total MS1 area for that transition was divided by the normalization peptide sum for that experiment and this normalized ratio was taken as the value for each 471 Da modified transition. Full data plus Limelight links can be found in Supplementary File 2, Sheet 3. Full Skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url.

Supplementary Note 1: Magnum

Overview

Magnum is a database search algorithm designed to identify adducts of variable mass without *a priori* knowledge. The algorithm has three major functional categories: reading and processing inputs, open modification database searching, and results scoring (Figure S1). Output for Magnum is a simple tab-delimited text format. The output is optionally also exported in PepXML[10] format, for potential use with existing software supporting this format. Magnum is written in C++, and is open source and freely available from http://magnum-ms.org/.

The basic software framework and analytical functions for Magnum are adapted from the cross-link identification software Kojak[11], and the data processing in many instances are identical. For example, the data structures, spectral processing, and Xcorr scoring algorithm (originally adapted from Comet[4,12]) remain the same. Magnum, however, is entirely distinct from Kojak in that Magnum cannot perform cross-linked spectra searching and Kojak cannot perform open modification searching. The principal searching function of each algorithm is mutually exclusive, and neither algorithm can complete the task of the other. Thus, as the many of the inner workings of Magnum have already been described[11], and its scoring scheme well known throughout proteomics[4,12,13], we focus particularly on the novel aspects that were developed for open modification searches.

Reading and Processing Inputs

Magnum requires a protein FASTA sequence file and a spectral data file. The FASTA sequence file is parsed to create a peptide list according to the user-defined proteolytic cleavage rules (see Table S1). Peptides containing suspected adduct attachment sites are marked to facilitate downstream analysis of these sequences. The user can also alter the mass of any amino acid, specify novel characters for special-case amino acids, and identify both static and variable modification masses on either amino acids, peptide termini, or protein termini. Decoy protein sequences can be identified with a user-defined text label, and peptides mapping to these protein sequences will be identified as decoy PSMs in the Magnum results, for the purposes of downstream validation at the user's discretion.

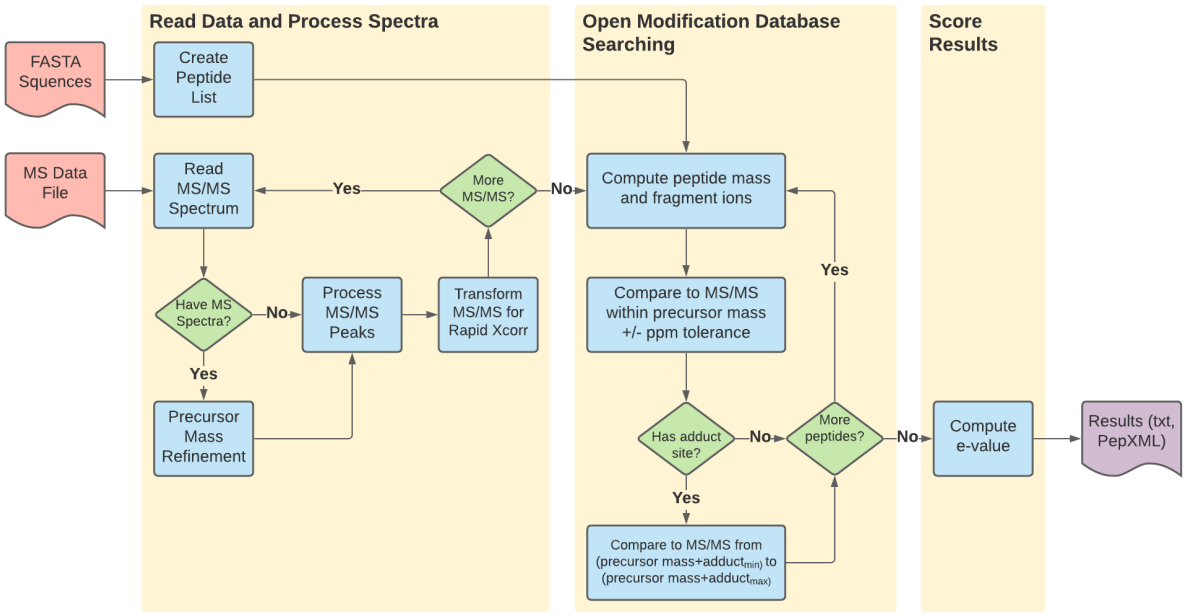


Figure S1: Magnum data processing workflow. Magnum architecture is divided into three primary categories: reading and processing spectra, open modification database searching, and scoring. Input files are a protein FASTA sequences file and spectral data in an open format (e.g. mzML). Output from Magnum is in both tab-delimited text and PepXML format.

Table S1: Select parameters relevant to open modification searches in Magnum

Open Modification Search Parameters

Parameter	Values	Description
adduct_sites	Uppercase [A-Z],n,c	Identifies one or more sites of open modifications. Must be uppercase letters, except for lowercase 'c' and 'n' to indicate protein C-terminus and protein N-terminus.
min_adduct_mass	number	Describes the smallest adduct mass in the open modification search range.
max_adduct_mass	number > min_adduct_mass	Describes the largest adduct mass in the open modification search range.

Peptide Sequence Search Rules

Parameter	Values	Description
enzyme	enzyme cleavage rules	Specifies the peptide sequence cleavage rules to produce peptides. Rules are defined as amino acids where cleavage occurs. See http://magnum-ms.org/param/enzyme.html
max_miscleavages	positive number	Maximum number of missed enzyme cleavages to consider when computing the peptide list from the FASTA sequence file.
min_peptide_mass	number	Smallest peptide mass allowed in the search space (including the adduct mass).
max_peptide_mass	number > min_peptide_mass	Largest peptide mass allowed in the search space (including the adduct mass).
min_peptide_length	positive number	Minimum number of amino acids (regardless of mass) in any peptide to search.
max_peptide_length	number > min_peptide_length	Maximum number of amino acids (regardless of mass) in any peptide to search.

Select Parameters to Customize Search Results

Parameter	Values	Description
modification	amino acid, modification mass	Indicates site and mass for variable modifications of known mass. This parameter can be repeated any number of times, each specifying a novel variable modification to search.
max_mods_per_peptide	positive number	Maximum number of variable modifications to consider on a peptide. The larger the number, the slower Magnum performs.
fixed_modification	amino acid, modification mass	Indicates fixed amino acid modification mass applied to all instances of the amino acid. Example: "C 57.02146" for carbamidomethyl-cysteine.
isotope_error	positive number	Integer value indicating number of carbon atom offsets to consider when evaluating precursor mass predictions.
e_value_depth	positive number	Number of decoy peptides per histogram when computing e-values for each PSM. Recommended to have at least 5000. However, larger numbers increase computation time for Magnum.
split_percolator	1 or 0	Creates two Percolator input files, one for peptides without open modifications, and the other for peptides with open modifications. Each file can be given to Percolator individually, to calculate FDR independently on each type of result. Setting the value to 1 activates this feature. Setting it to 0 deactivates this feature. By default, this feature is turned off.

The spectral data input file must be one of several supported open formats that include mzML (preferred), mzXML, and MGF. Magnum reads the spectral data file to extract MS/MS scan data, performs refinement steps, and converts the spectra to an internal data structure for rapid cross-correlation analysis[11,14] before storing all spectral data in memory. Refinement consists of two major processes. The first process performs analysis of precursor MS spectra to refine the precursor mass. Precursor refinement attempts to find the elution apex of the selected peptide represented in an MS/MS spectrum, to more accurately predict the monoisotopic precursor mass and charge state, particularly if such information is not provided in the spectral data file. Precursor refinement is skipped if the data contain no precursor MS spectra. Additional functions allow for estimation of isotope mass errors and additional charge state assignments among ambiguous or missing precursor information. For many spectra analyzed by Magnum, more than one candidate precursor mass and charge state may be assigned for database searching. The second major refinement process consists of MS/MS peak refinement. Here, isotope clusters are collapsed to their monoisotopic peak, summing the intensities of each peak in the cluster. Subsequent fragment ion matching (see below) is therefore performed on the monoisotopic mass. Additional, optional processing includes reducing the MS/MS spectra to a fixed, user-defined maximum number of peaks. These steps are repeated on all MS/MS spectra before proceeding to the database searching procedures.

Open Modification Database Searching

Magnum attempts to identify peptide sequences from MS/MS spectra allowing for an open modification mass with a user-defined range, referred to as an adduct mass. Database searching is performed by matching theoretical fragment ions (*a*, *b*, *c*, *x*, *y*, or *z*, user-defined) for every peptide sequence parsed from the FASTA file to every spectrum for which the peptide mass falls within the precursor mass tolerance. Which spectra fall within this mass range is defined as:

Equation S1: $$p + v + a_{min} \leq s_{pre} \leq p + v + a_{max}$$

Where p is the peptide mass, v is the sum of the variable modification masses (if any), and a_{min} and a_{max} are the smallest and largest adduct masses defined by the user. s_{pre} is a precursor mass assigned to an MS/MS spectrum. For peptides without an adduct binding site, $a_{min} = a_{max} = 0$, defining a narrow mass range with a user-defined ppm tolerance around the precursor mass for spectra to search. For peptides with an adduct site, the mass range may span hundreds of daltons and require searching several thousand spectra. The adduct mass is different for each spectrum, and defined as:

Equation S2: $$a_{pre} = s_{pre} - p - v$$

Where a is the adduct mass for a predicted precursor (*pre*) of spectrum s, and p and v are the peptide and variable modification masses, respectively. If the peptide contains more than one adduct site, the adduct mass is iteratively scored at each site and the highest scoring orientation is kept. In this manner, it is possible to localize the adduct on the peptide, however, no probability

is assigned. Thus, the localization is simply the highest scoring orientation without validating the likeliness that this orientation is correct given all available options. The adduct mass is never divided among multiple sites, and therefore only a single adduct of variable mass is ever scored per peptide.

Results Scoring

Peptide-spectrum matches (PSMs) are initially scored using a cross-correlation scoring method (Xcorr) as previously implemented in Comet[4,12] and Kojak[11]. Briefly, theoretical fragment ions of equal weight are matched to observed, locally normalized spectral peaks within a mass tolerance bin the user can adjust to reflect the resolution of their instrument[15]. However, Xcorr values are not generally comparable between PSMs, as the Xcorr values for longer peptide sequences tend to be higher than Xcorr values for shorter peptide sequences. This is because longer peptide sequences contain more theoretical fragment ion masses to match to a MS/MS spectrum, giving them a higher potential score. A solution to this problem is to compute an expect value (e-value) for the PSM with the highest Xcorr value for each spectrum[4], using the histogram of all PSM Xcorr values to that spectrum. This generally works well if the assumption that all PSMs for a given spectrum are approximately the same length. However, for open modification searches, the PSMs for a spectrum have a much larger range of peptide lengths when considering a short peptide with a large adduct vs. a long peptide without any adduct. Therefore, an alternative method was used to compute e-values for all PSMs for a spectrum (not just the highest Xcorr value), then re-rank the PSMs by e-value and return the PSM with the lowest e-value to the user. The lowest e-value PSM is returned from all modified and unmodified forms of peptide sequences that fall within the search constraints of the spectrum being assessed.

To more accurately compute an e-value from an Xcorr value for a PSM of a given length, a histogram of Xcorr values of random peptide sequences of equal length was generated for each spectrum. The random peptides are selected from approximately 350,000 amino acids worth of *Drosophila melanogaster* protein sequences, a technique adapted from the Comet algorithm, and also representative of real peptide sequences. Specifically, an array of 5,000 peptide sequences of 70 amino acids in length, and their respective b- and y-ion neutral mass values at every amino acid breakpoint along the peptide sequences, is precomputed and stored in memory. For each spectrum, these neutral mass values are compared to the spectrum, up to the point at which either a) the desired peptide length is achieved, or b) the precursor mass is exceeded. Should the user request alternative ions (such as c- or z- ions for ETD), a simple arithmetic adjustment is made to the neutral mass fragment ions (e.g. adding 17.0265 Da to the neutral mass value of a b-ion to produce the mass value of a c-ion). If the random peptide was less than the expected mass, an adduct mass was added to the fragment ions. The placement of the adduct was at the N-terminus, C-terminus, or middle of the random peptide, resulting in three independent peptide measurements for each random peptide. Therefore, for each spectrum, a set of histograms was generated for each peptide length from these random peptides, and accounting for an adduct mass, and accounting for different localizations of that adduct mass, representing the expected distribution of random peptides at any given length for that spectrum. From these distributions, an e-value could be computed for any peptide score by using the histogram of equivalent peptide length. Thus, the e-value for a short peptide of 10 amino acids with a large adduct was generated using a histogram of Xcorr values from random peptides of 10 amino acids in length. For the same spectrum, the e-value for a different peptide of 14 amino acids with a small adduct is then computed using a histogram of Xcorr values from random peptides of 14 amino acids in length. By using this approach, the 10-amino acid PSM may produce a lower e-value than the 14-amino

acid PSM, despite having a lower Xcorr value. The e-values normalize the effects of peptide length represented by Xcorr values, allowing peptides of very disparate lengths to be compared for the same spectrum. This step of the results scoring is very computationally intensive, which is mitigated by pre-computing the histograms prior to the database search. The pre-computation process is made efficient by taking all the histogram Xcorr values for peptides of length *n*, and extending them by one additional fragment ion to produce histograms of length *n+1*. This process is repeated for the range of all expected peptide lengths for a given MS/MS spectrum after considering all possible adduct sizes.

MS-labile versus stable adducts

MS-labile modifications are prone to dissociation during peptide fragmentation preceding MS/MS acquisition and may result in a strong unmodified ion series. In contrast, stable adducts remain attached to the peptide through the fragmentation process and result in an adduct-modified ion series. Both these situations are included in the Magnum search space consisting of every peptide sequence considered for a given spectrum. Each possible outcome of that entire search space (each candidate peptide sequence, without modification or with modification, and that modification localized or not) is scored for each spectrum and the highest scoring PSM returned as the result. Peptide sequence and adduct mass identification is therefore possible with or without adduct localization.

Restriction of open modification mass to specific residues in Magnum

Magnum optionally allows adduct localization to be restricted to specific amino acids. The reactivity of xenobiotics may be known or hypothesized based on the chemistry of the compound of interest, or detection of glutathione or other conjugates to reactive intermediates. Increased search sensitivity and statistical power can be gained by restricting the open modification search space to specific amino acids[16].

Adduct reporter ions

We implemented the ability of Magnum to flag peptide spectrum matches (PSMs) that contained user-defined reporter ion masses and of Limelight to filter data using this information and to annotate reporter ions when viewing individual spectra using Lorikeet[17]. This is useful to identify MS/MS spectra that contain peptides modified with labile adducts whose fragment masses will be independent of the peptides to which they are adducted. Details of this parameter can be found on the Magnum website here.

Error estimation and PSM utilities in Magnum

Differences in unmodified and modified peptide search spaces and the small proportion of PSMs representing specific xenobiotic adducts of interest make accurate error estimation difficult in open searching[18].

Magnum contains a parameter (split_percolator) to divide PSMs into separate classifications of those that contain an open modification, and those PSMs that do not: http://magnum-ms.org/param/split_percolator.html It is possible then to perform Percolator validation on unmodified and open-mass modified PSMs separately with the aim of improving the sensitivity of the analysis.

We performed an analysis of the use of this function using the gold standard methods and data described in Figure 1 and Supplemental Note 2, however we found no improvement in the already excellent accuracy and sensitivity of identifying xenobiotic-protein adducts by using this option

(the data points for precision and recall for our gold standard spectra were no different between the two analyses). This result is explained by the fact that in open searching, the search space is dominated by open-mass modified peptides (Figure S2). When a target-decoy approach is used for error estimation (e.g. as implemented in Percolator) the decoy identifications are almost exclusively to open-mass modified decoys due to the proportionally vast size of the open-mass modified search space, while target peptide identifications are split between the unadducted peptides prevalent in biological samples and the adducted (open-modified) peptides.

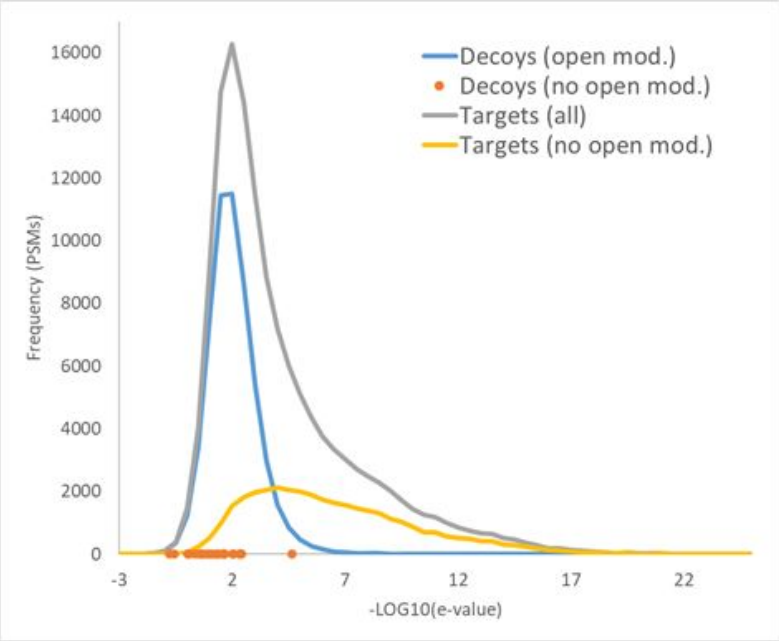


Figure S2: Distribution of –log base 10 e-values of targets and decoys from a single Magnum example search. The blue line corresponds to decoys with an open modification mass, the orange points correspond to decoys without an open modification mass (there are 27 points). The grey line corresponds to all target PSMs and the yellow line corresponds to unmodified targets only.

Given that there are so few unmodified decoys in comparison to modified decoys, training Percolator on unmodified data separately does not result in enough decoy data to discriminate unmodified decoys from unmodified targets, while there is sufficient representation of open-modified targets as well as open-modified decoys for discrimination.

To further clarify this scenario, if the search was done without open modifications there would be many unmodified target PSMs and many unmodified decoy PSMs. If the search is done with open modifications, all the previous top unmodified decoy and target candidates are still present; however, the search space is now greatly expanded with thousands-fold more open modification candidates. Because of this vastly expanded search space, open searching results in many more target PSMs (now both modified and unmodified) and nearly exclusively modified decoy PSMs, since previous top-scoring unmodified candidates are now largely outcompeted by the much larger number of modified candidates by chance

To determine if the score histograms of open-modified and unmodified peptides largely coincide with each other, we combined 24 separate searches against the same background FASTA file to generate enough unmodified decoy PSMs to plot representative distributions. The resulting

histogram (Figure S3) shows that the distributions of the decoy PSMs with and without open-mass modifications are similar, with the unmodified PSMs slightly shifted towards worse e-values.

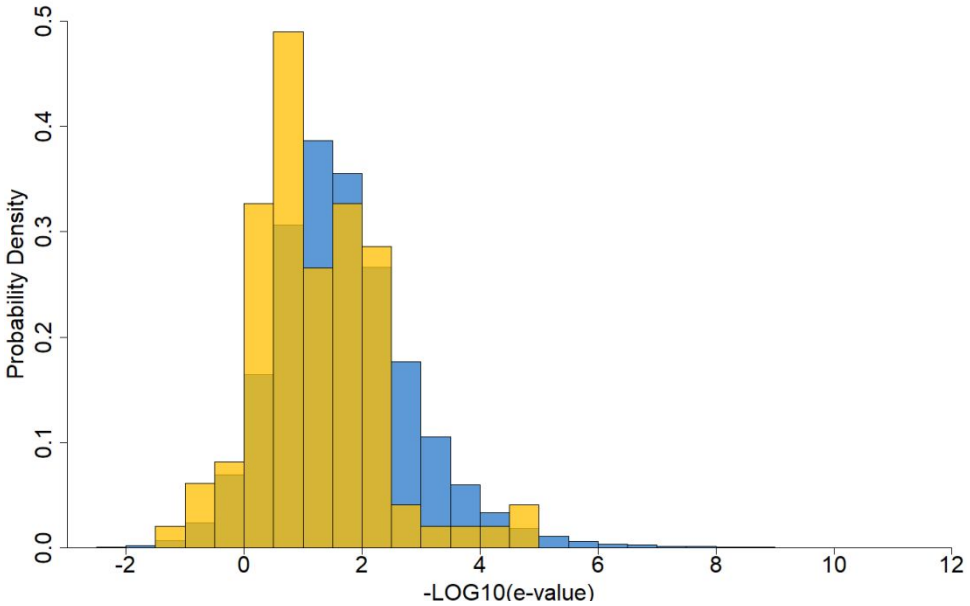


Figure S3: Histogram of –log base 10 e-values of decoys from Magnum from 24 example searches. The blue bars correspond to decoys with an open modification mass, the yellow bars correspond to decoys without an open modification mass. The number of points for the blue bars is 462,835. The number of points for the yellow bars is 98.

As such the modified decoy distribution can be used to find an error rate for the unmodified target distribution although this likely sets up an overly conservative FDR cutoff rate for unmodified targets. Yet, as our work is focused on identification of open-modified PSMs we do not consider this to be a problem in our data.

The many types of error rate analyses possible in open searching and their resulting improvements in sensitivity are beyond the scope of this manuscript, and were not considered in our analytical methodology, however Limelight was designed to support output from any software pipeline, and we look forward to supporting the future improvements in this very active field as they are developed.

Magnum search engine speed and search space recommendations

Magnum can perform searches against entire proteomes and will typically complete such searches on a desktop computer within a few hours. This is not our recommended method of searching for low abundance xenobiotic-protein modifications. For this we favor using a standard closed search (e.g. comet) to identify the subset of proteins present in a sample and then searching this smaller database in open search mode. This improves statistical power[16], decreases search times, and reduces the complexity of subsequent data analysis. This is the method we used in the current manuscript for all samples except for the phospho-protein benchmarking, which was searched against the entire human proteome. Our method of creating a sub-database is described in Materials and Methods above.

If adduct sites are known as is the case for many drugs, we recommend restricting the allowed adduct sites in the Magnum configuration using the (adduct_sites) parameter: http://magnum-

ms.org/param/adduct_sites.html. If modified amino acids are unknown all residues can be entered as adduct sites.

Supplementary Note 2: Gold standard dataset for evaluation of xenobiotic-protein adduct discovery

As the focus of the current study was to accurately detect unknown xenobiotic-protein adducts related to specific exposures, the assignment of correct open modification masses is critical. We therefore created three gold standard datasets to allow evaluation of both the accuracy (precision) and sensitivity (recall) of, Magnum, the search algorithm presented here, as well as previously published open search algorithms, within the context of our xenobiotic-protein adduct discovery pipeline. These datasets were derived from four dicloxacillin and flucloxacillin treated HSA samples and the raw MS data files (Table S2) were deposited to the ProteomeXchange Consortium via the PRIDE[19] partner repository with the dataset identifier PXD025019. These data consist of 307,652 MS/MS scans. We derived known correct open modification masses for 2,979 unique MS/MS spectra from these data using two methods.

Table S2: Raw files used for creation of gold standard data.

Sample Treatment	Untreated (rep 1)	Untreated (rep 2)	Flucloxacillin (rep 1)	Flucloxacillin (rep 2)	Dicloxacillin (rep 1)	Dicloxacillin (rep 2)
Filename	QEP2_2018_0812_AZ_024_az732_AZ.mzML	QEP2_2018_0812_AZ_025_az733_AZ.mzML	QEP2_2018_0812_AZ_028_az734_AZ.mzML	QEP2_2018_0812_AZ_029_az735_AZ.mzML	QEP2_2018_0812_AZ_029_az735_AZ.mzML	QEP2_2018_0812_AZ_033_az736_AZ.mzML
Used in gold standard dataset	no	no	yes	yes	yes	yes

Both methods made use of the fact that β-lactam antibiotics and their adducts fragment in MS/MS giving rise to known reporter ions[20,21] (Figure S4, green boxes).

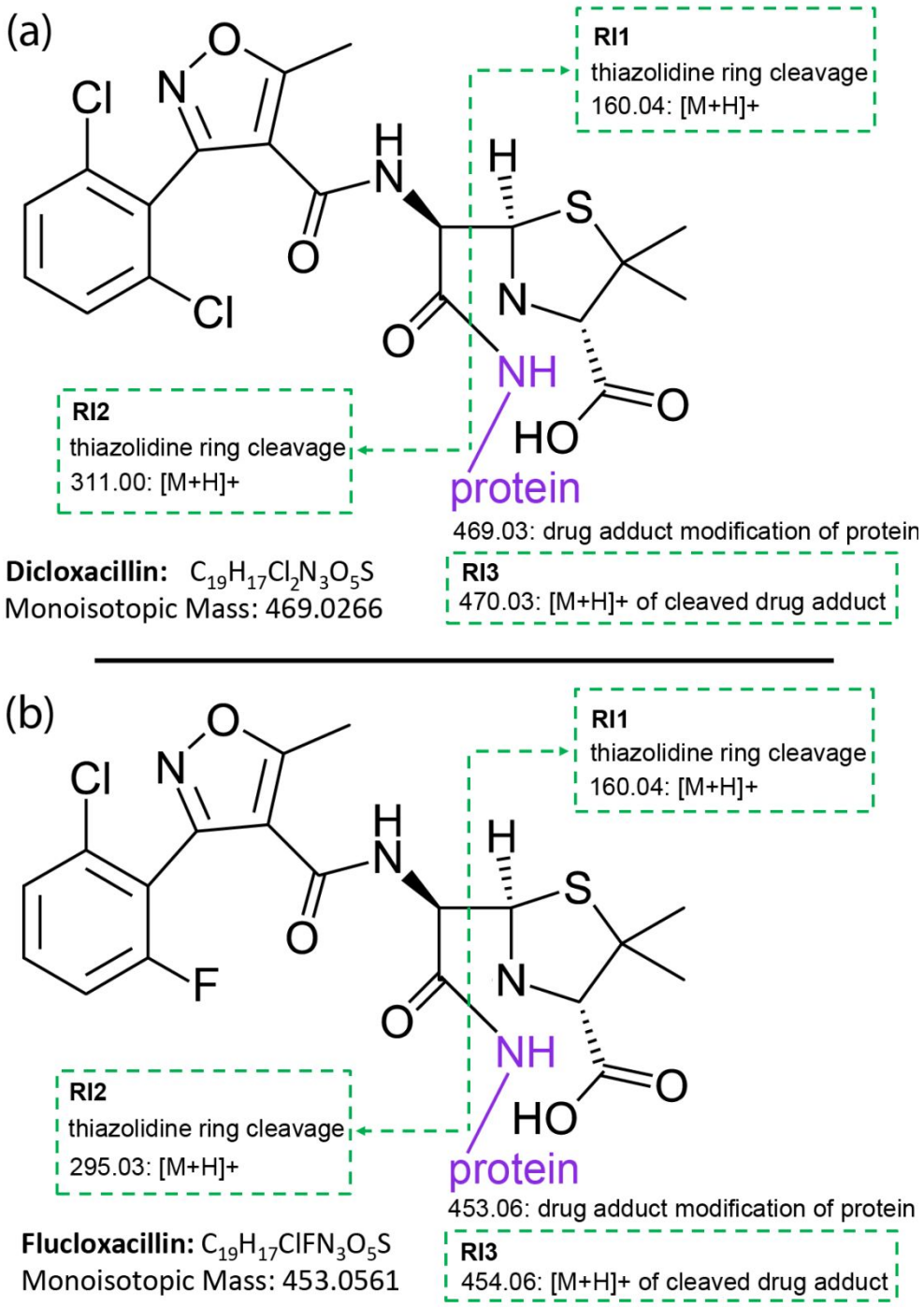


Figure S4: The structure of (a) dicloxacillin and (b) flucloxacillin after forming adducts on a lysine primary amine. The monoisotopic mass and chemical formular of the original antibiotics are listed. For both antibiotics, the adduct is covalently bound to the lysine primary amine (purple). Dicloxacillin and flucloxacillin form MS-labile adducts. During peptide fragmentation the adduct itself is cleaved at the thiazolidine ring (green dotted line) releasing reporter ion 1 (RI1) and reporter ion 2 (RI2). The entire adduct can also be cleaved from the peptide releasing reporter ion 3 (RI3). The masses of these ions are independent of the peptide to which the drug is adducted as the ions are derived from adduct fragmentation. The adduct masses and those of the reporter ions have been previously characterized[20,21].

These reporter ions are present within any MS/MS spectrum that contains a β-lactam antibiotic adduct, and their masses are unrelated to the peptide to which the adduct is attached. We wrote a simple program called ScanFinder, available at http://magnum-ms.org, which searches raw spectra for signals at a defined m/z and intensity (percent of the base peak). Raw MS data acquired from untreated, flucloxacillin and dicloxacillin treated HSA were searched for both dicloxacillin and flucloxacillin reporter ions at the specified intensities. The results of these searches are summarized in Table S3 and show that zero spectra from untreated HSA samples contained the combination of reporter ions searched for at the required intensities. Likewise, zero spectra in flucloxacillin treated samples contained dicloxacillin reporter ions and zero dicloxacillin treated samples contained flucloxacillin reporter ions at the specified intensities. This is important as it shows that spectra picked out in the treated samples are specific to their respective treatments.

Table S3: Number of MS/MS spectra that contain β-lactam antibiotic reporter ions at the defined m/z and intensities.

Sample Treatment	Untreated (rep 1)	Untreated (rep 2)	Flucloxacillin (rep 1)	Flucloxacillin (rep 2)	Dicloxacillin (rep 1)	Dicloxacillin (rep 2)
Number of MS/MS spectra containing dicloxacillin reporter ions 160.04 m/z ≥ 80% intensity 311.00 m/z ≥ 5% intensity 470.03 m/z ≥ 5% intensity	0	0	0	0	654	587
Number of MS/MS spectra containing flucloxacillin reporter ions 160.04 m/z ≥ 80% intensity 311.00 m/z ≥ 5% intensity 470.03 m/z ≥ 5% intensity	0	0	798	758	0	0

In Gold Standard Method 1, the 1,241 MS/MS spectra from the dicloxacillin treated HSA samples that were found to contain dicloxacillin specific reporter ions at the intensity threshold of 80% 160.04 m/z, 5% 311.00 m/z and 5% 470.03 m/z, were evaluated to confirm the presence of a 469 Da adduct modification in the spectrum. As there were too many spectra to manually solve each of the 1,241 spectra individually, the following method was used to create a list of scan numbers representing spectra resulting from a peptide with a single 469 Da mass modification:

- 1,241 MS/MS spectra were found to contain dicloxacillin specific reporter ions with a minimum of the following intensities: 80% 160.04 m/z, 5% 311.00 m/z and 5% 470.03 m/z
- The targeted m/z of each MS/MS scan's precursor ion was noted and rounded down to the nearest integer yielding 145 distinct targeted m/z's.
- MS/MS spectra were extracted from the original list of 1,241 only for targeted m/z's that occurred at least 10 times resulting in 926 scans with 33 distinct m/z's. This constituted 75% of the original spectra (946/1,241).
- For each of the 33 distinct sets of targeted precursor ions one representative spectrum was manually evaluated to confirm the peptide contained a single 469 Da mass modification. If this could be confirmed, the entire set of scans at that targeted m/z was added to the gold standard list, if this could not be confirmed the entire set was excluded from the list.

Using this completely algorithm free method, Gold Standard Method 1 resulted in 763 MS/MS spectra of dicloxacillin treated HSA that were confirmed to contain a single 469 Da dicloxacillin adduct modification. These scans represented 22 distinct m/z species (Figure S5a).

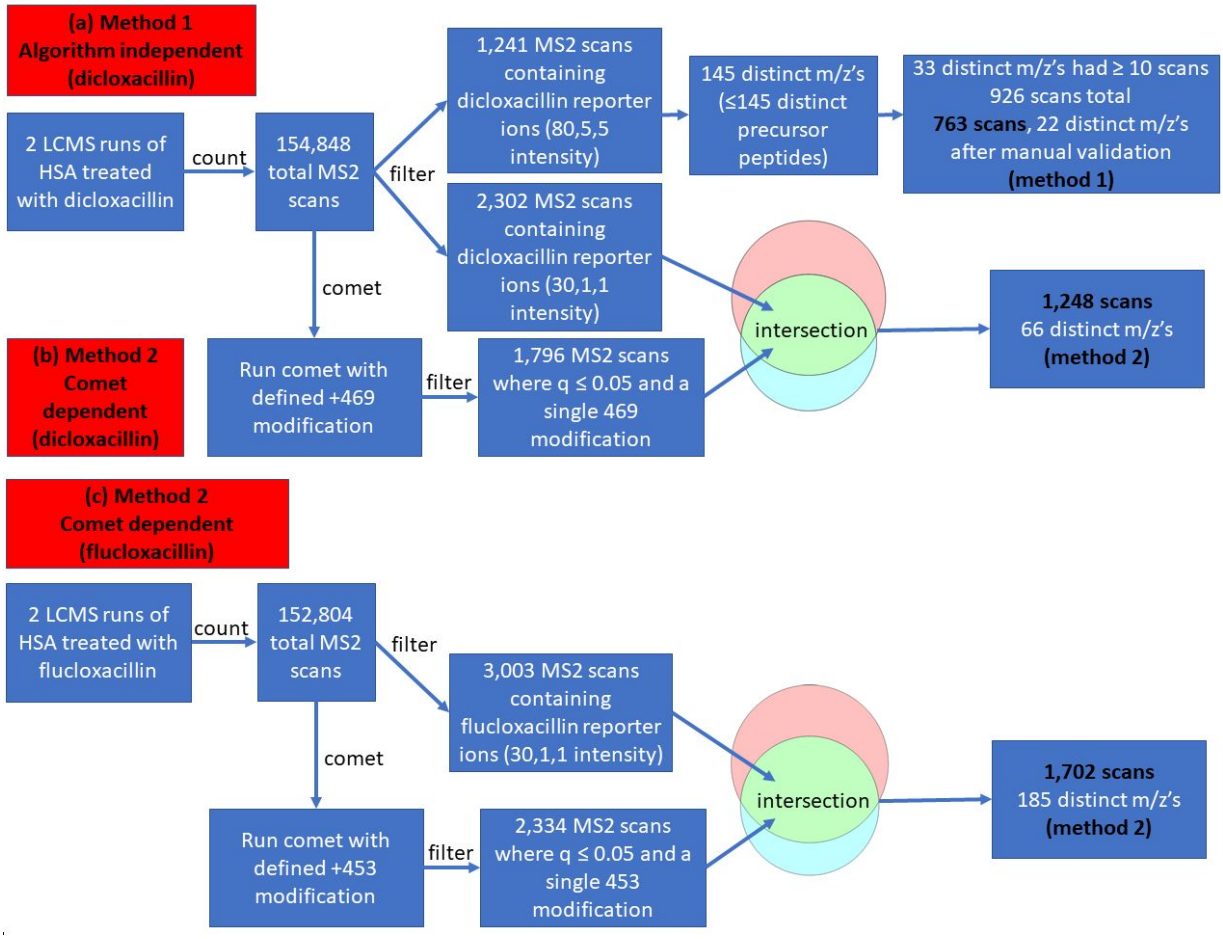


Figure S5: Workflow for the creation of 3 gold standard datasets. (a) Method 1 was fully manual, relied on no algorithms and resulted in 763 MS/MS scans with a known 469 Da dicloxacillin modification. (b,c) Method 2 relied on the presence of known β-lactam antibiotic reporter ions plus confident adduct modification mass identifications using a comet closed search. This method resulted in 1,248 and 1,702 MS/MS scans with known dicloxacillin or flucloxacillin adduct modifications, respectively.

In Gold Standard Method 2, each of the 6 raw files described in Table S2 were searched using ScanFinder for dicloxacillin and flucloxacillin reporter ions similarly as for method 1, but using the lower intensities stated in Table S4.

Table S4: Number of MS/MS spectra that contain β-lactam antibiotic reporter ions at the defined m/z and lower intensities than searched for in Table S3.

Sample Treatment	Untreated (rep 1)	Untreated (rep 2)	Flucloxacillin (rep 1)	Flucloxacillin (rep 2)	Dicloxacillin (rep 1)	Dicloxacillin (rep 2)
Dicloxacillin reporter ions 160.04 m/z ≥ 30% intensity 311.00 m/z ≥ 1% intensity 470.03 m/z ≥ 1% intensity	0	0	0	0	1217	1085
Flucloxacillin reporter ions 160.04 m/z ≥ 30% intensity 311.00 m/z ≥ 1% intensity 470.03 m/z ≥ 1% intensity	0	0	1526	1477	13	5

The ScanFinder results in Table S4 show that zero spectra from untreated HSA samples contained the combination of reporter ions searched for at the required lower intensities. Likewise, zero spectra in flucloxacillin treated samples contained dicloxacillin reporter ions at the specified lower intensities. 18 out of 154,848 spectra from dicloxacillin treated samples contained flucloxacillin specific reporter ions at the lower intensities used for this second set of ScanFinder searches. This is likely due to carryover from the flucloxacillin samples, which were run before the dicloxacillin samples on the same column. These data thus show that spectra picked out in the treated samples are specific to their respective treatments even at these lower intensity thresholds.

In addition to searching for MS/MS scans containing dicloxacillin and flucloxacillin reporter ions, we ran a closed comet search on each dataset, allowing for a defined variable modification of 469.0266 (the known dicloxacillin adduct mass) or 453.0561 (the known flucloxacillin adduct mass) on lysines. Scan numbers resulting in a confident (Percolator assigned q ≤ 0.05) PSM containing either a single dicloxacillin or flucloxacillin adduct were noted (Table S5).

Table S5: Number of MS/MS spectra that contain a single defined 469.0266 (dicloxacillin) or 453.0561 (flucloxacillin) adduct at a Percolator q ≤ 0.05 based on a comet closed search.

Sample Treatment	Untreated (rep 1)	Untreated (rep 2)	Flucloxacillin (rep 1)	Flucloxacillin (rep 2)	Dicloxacillin (rep 1)	Dicloxacillin (rep 2)
Spectra containing a single 469.0266 modification (comet)	140	119	124	184	916	880
Spectra containing a single 453.0561 modification (comet)	84	91	1149	1185	188	224

To create the Method 2 Gold Standard datasets we listed the scan numbers for the lower intensity ScanFinder search plus the scan numbers for the comet search specific to each antibiotic and extracted the intersection of those scan numbers. In other words, only scan numbers that showed the required reporter ions AND resulted in a confident comet PSM yielding a single antibiotic adduct of the specified mass were added to the gold standard list (Figure S4b and S4c). **Gold Standard Method 2 resulted in 1,248 MS/MS spectra (66 distinct m/z's) of dicloxacillin treated HSA confirmed to contain a single 469 Da dicloxacillin adduct modification. The same method yielded 1,702 MS/MS spectra (185 distinct m/z's) of flucloxacillin treated HSA confirmed to contain a single 453 Da flucloxacillin adduct modification.**

The procedure above resulted in 3 sets of gold standard scan numbers: (1) dicloxacillin gold standard method 1 (763 spectra with known +469 Da single modifications); (2) dicloxacillin gold standard method 2 (1,248 spectra with known +469 Da single modifications); and (3) flucloxacillin gold standard method 2 (1,702 spectra with known +453 Da single modifications). These 3 sets of scan numbers constituted a combined total of 2,979 unique MS/MS spectra all of which were modified by a single xenobiotic-protein adduct of known mass. Various searches were then run on the complete datasets and only the open modification masses for these specific spectra were extracted for the gold standard comparisons. Precision and recall could then be calculated for each method for any search performed, based on the fraction of correct answers and the total number of correct answers.

The data presented in the main manuscript (Main manuscript Figure 1) plus Figures S6 and S7 below, combine results from all three methods for brevity however results from the individual methods are also presented (Figures S8, S9, S10 and S11), illustrating that PSMs derived from each individual method produced similar results.

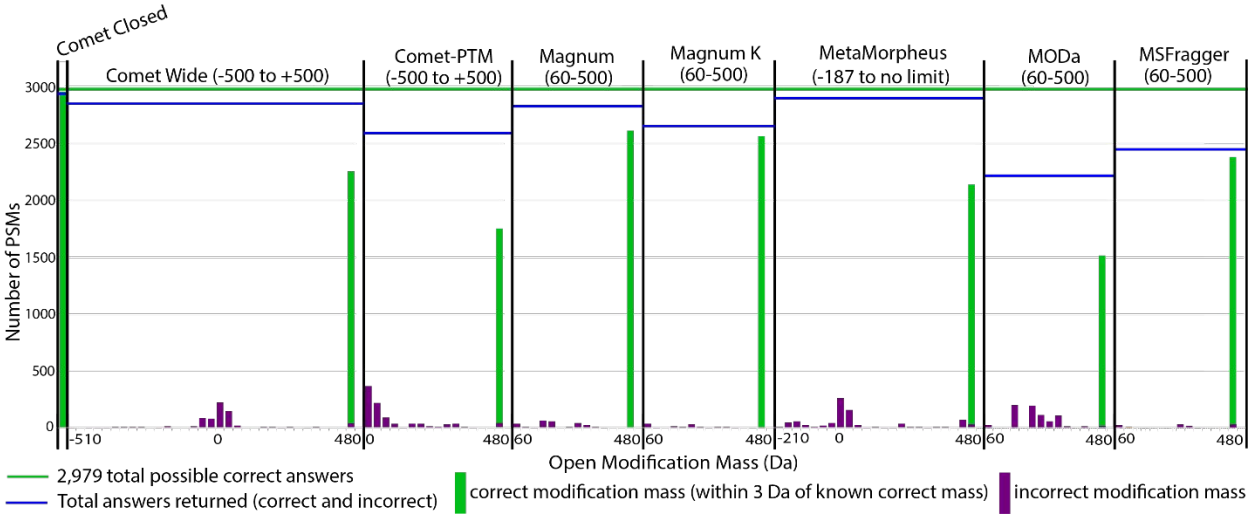


Figure S6: Histograms showing the distribution of open modification masses reported by 7 open search algorithms for 2,979 unique MS/MS spectra definitively determined to result from a peptide containing a single +469 Da (dicloxacillin) or +453 Da (flucloxacillin) modification using both methods 1 and 2 described above. Results are shown at 1% FDR and are distributed between incorrect masses (purple bars, >3 Da from known correct modification mass) and correct masses (green bars, within ±3 Da of the known correct modification mass). The first and last mass bin are labeled on the x axis. The open mass range searched by each algorithm is shown in parenthesis above each plot. Bins are 30 Da wide and all correct answers fall within the 450 Da bin. Incorrect masses within the 450 Da bin are shaded purple.

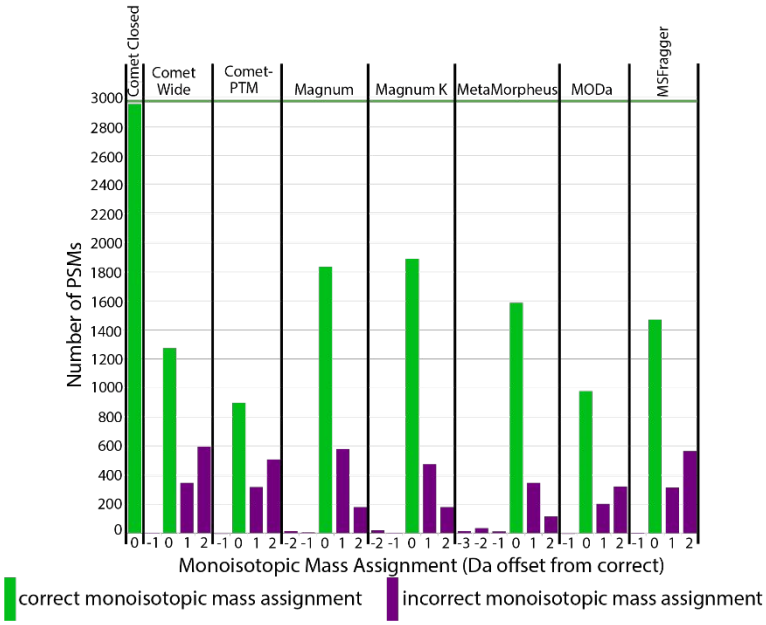


Figure S7: Histograms of monoisotopic masses assigned to each of the correct answers in Figure 1 and S6 by each algorithm (open-pFind is excluded due to sparsity of results). Results are distributed between incorrect monoisotopic mass assignments (non-zero offset values, purple bars) and correct monoisotopic mass assignments (zero offset values, green bars). Bin width is 1 Da. Correct monoisotopic mass assignments have 0 Da offsets. All data in this figure is filtered at a 1% FDR.

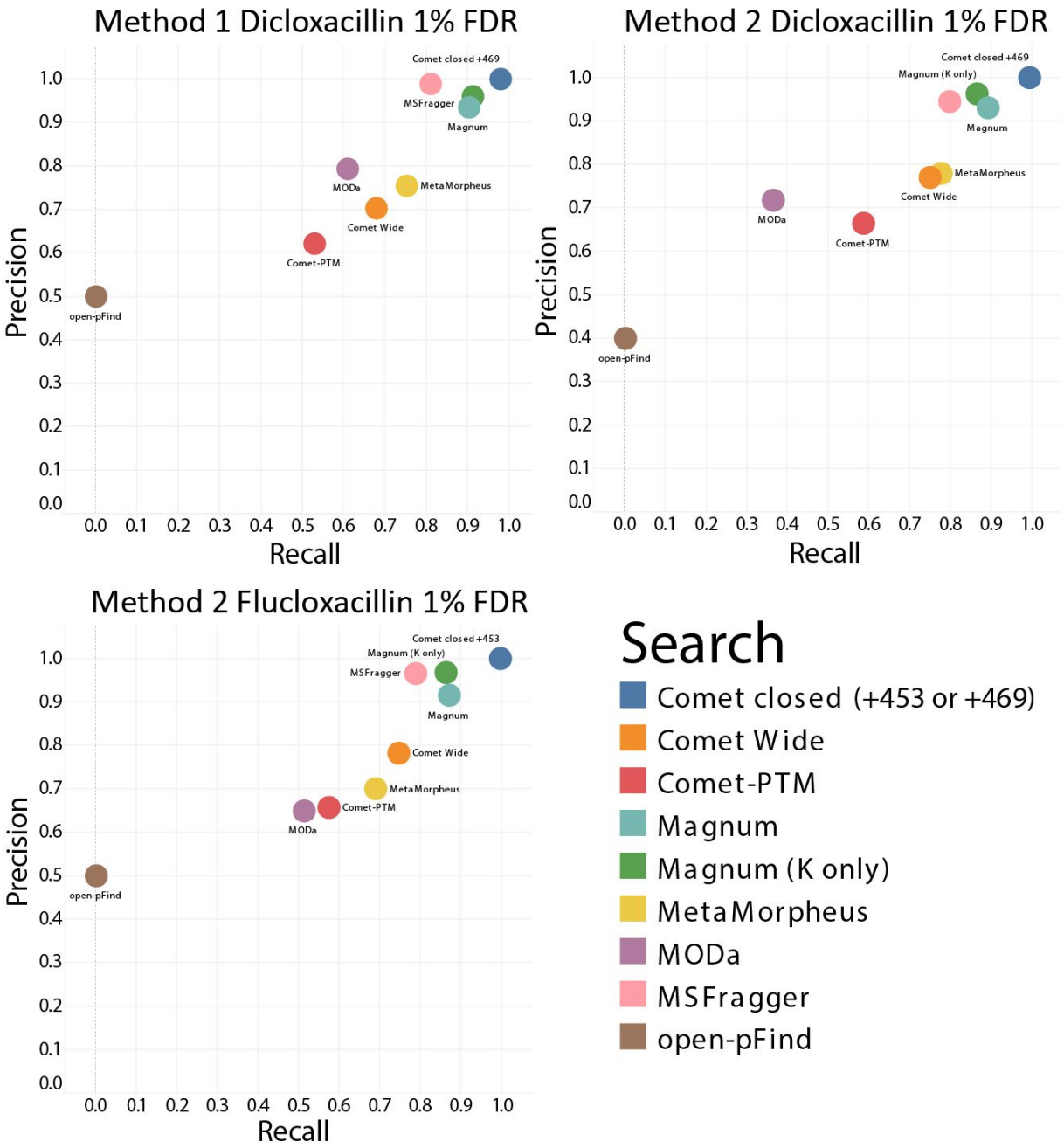


Figure S8: Precision recall plot of adduct masses reported at 1% FDR by 7 open search algorithms for MS/MS spectra definitively determined to result from a peptide containing a single +469 Da (dicloxacillin) or +453 Da (flucloxacillin) modification. Results shown in Main Manuscript Figure 1 are shown separately for each of the gold standards datasets described above. Results from closed comet searches using defined modifications of 469 or 453 were included as a positive control. An open modification mass returned by an algorithm is defined as correct if it is within ±3 Da of the known correct modification mass (469 or 453). Magnum was run allowing for open masses on any amino acid (Magnum) or restricted to lysines only (Magnum K), the previously published residue modified by dicloxacillin and flucloxacillin adducts.

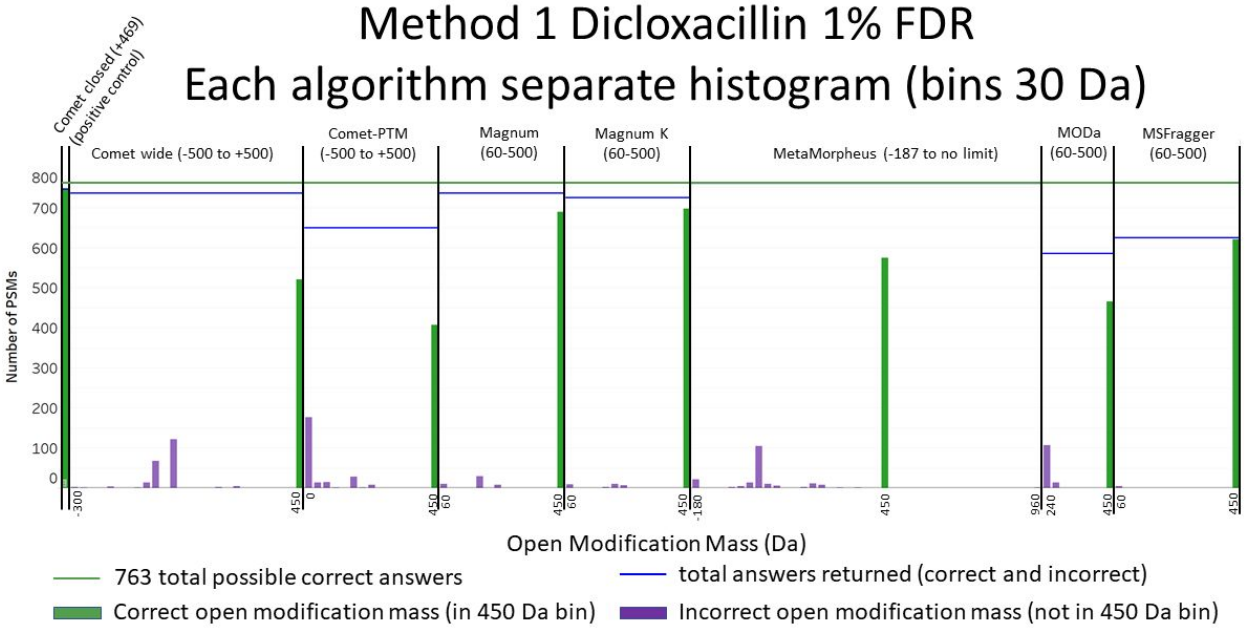


Figure S9: Histograms showing the distribution of open modification masses reported by each algorithm for the 763 gold standard spectra generated from dicloxacillin treated samples using method 1, described above. Results are shown at 1% FDR and are distributed between incorrect masses (purple bars, >3 Da from known correct modification mass) and correct masses (green bars, within ±3 Da of the known correct modification mass). The first and last mass bin are labeled on the x axis. The open mass range searched by each algorithm is shown in parenthesis above each plot. Bins are 30 Da wide and all correct answers fall within the 450 Da bin. Incorrect masses within the 450 Da bin are shaded purple.

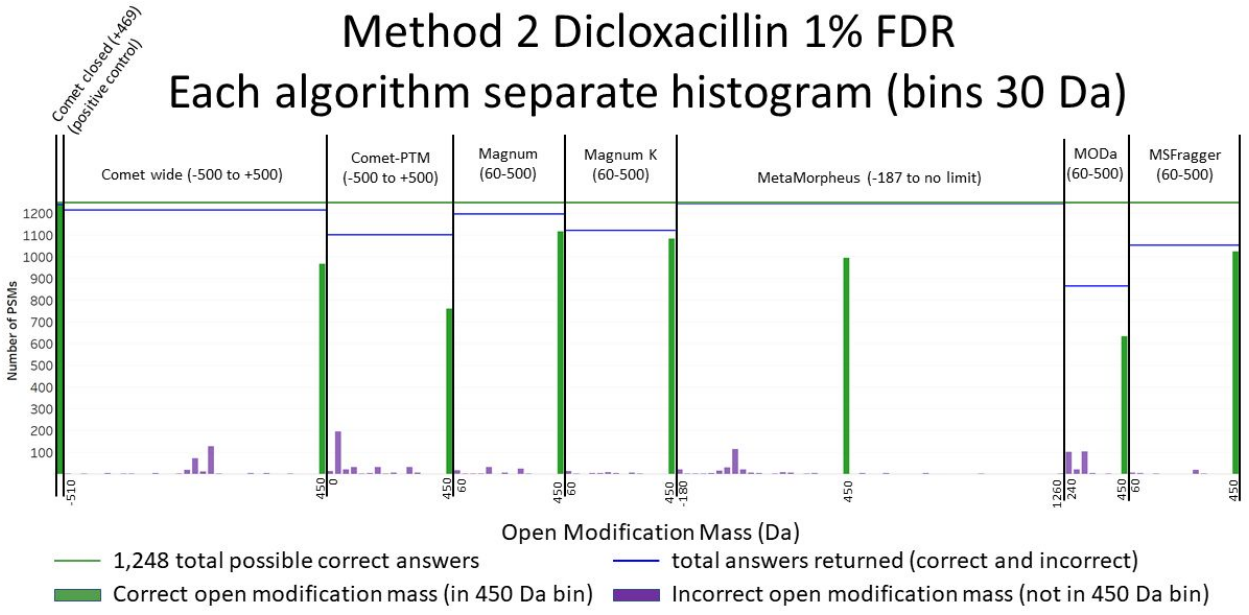


Figure S10: Histograms showing open modification masses reported by each algorithm for the 1,248 gold standard spectra generated from dicloxacillin treated samples by method 2, described above. Results at 1% FDR are distributed between incorrect masses (purple bars, >3 Da from known correct mass) and correct masses (green bars, within ±3 Da of the known correct modification mass). The first and last mass bin are labeled on the x axis. The open mass range searched by each algorithm is shown in parenthesis above each plot. Bins are 30 Da wide and all correct answers fall within the 450 Da bin. Incorrect masses within the 450 Da bin are shaded purple.

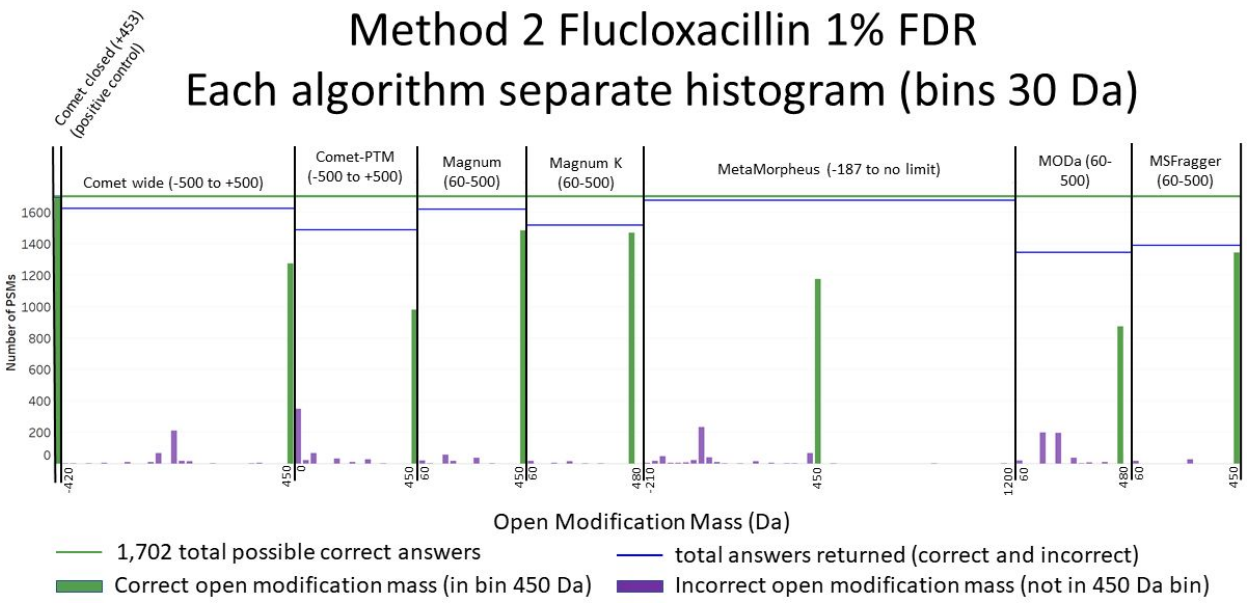


Figure S11: Histograms showing the distribution of open modification masses reported by each algorithm for the 1,702 gold standard spectra generated from flucloxacillin treated samples using method 2, described above. Results are shown at 1% FDR and are distributed between incorrect masses (purple bars, >3 Da from known correct modification mass) and correct masses (green bars, within ±3 Da of the known correct modification mass). The first and last mass bin are labeled on the x axis. The open mass range searched by each algorithm is shown in parenthesis above each plot. Bins are 30 Da wide and all correct answers fall within the 450 Da bin. Incorrect masses within the 450 Da bin are shaded purple.

Table S6: Gold standard results returned by open-pFind at 1% FDR. Of the 9 results returned, 4 had masses within the tolerance deemed correct (flucloxacillin 453 ± 3 Da; dicloxacillin 469 ± 3Da) for the purposes of our gold standard analysis. The masses matched by open-pFind correspond to the following three Unimod[22] entries: (1) Accession #: 409; Interim Name: FMN; Monoisotopic Mass: 454.088965. (2) Accession #: 1431; Interim Name: Hex(1)NeuAc(1); Monoisotopic Mass: 453.148240. (3) Accession #: 1375; Interim Name: dHex(1)Hex(2); Monoisotopic Mass: 470.163556. As open-pFind is unable to assign the real masses of flucloxacillin (453.0561) and dicloxacillin (469.0266) adducts, most spectra do not result in correct or confident assignments. This issue will exist for any modifications not in the Unimod database prior to searching.

URL to Annotated Spectrum on Limelight	Open Mod Mass	Reported Peptide	q-value
HSA + flucloxacillin - replicate 1			
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1893/psm/78163210	454.08897	YKAAFTEC[454.09]CQAADK	0.0031088
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1893/psm/78172798	38.930957	MPC[57.02]AED[37.95]YLSVVLN[0.98]QLCVLHEK	0.0073529
HSA + flucloxacillin - replicate 2			
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1894/psm/78198359	453.14824	RYKAAFT[453.15]ECC[57.02]QAADK	0.0010267
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1894/psm/78210881	324.03587	n[324.04]KASSAKQR	0.0036443
HSA + dicloxacillin - replicate 1			
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1895/psm/67218827	38.01565	DEGK[38.02]ASSAKQR	0.0011351
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1895/psm/67219788	0	AVM[15.99]DDFAAFVEK	0.012616
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1895/psm/67222933	470.16356	AT[470.16]EEQLK	0.0011351
HSA + dicloxacillin - replicate 2			
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1896/psm/71787832	37.946941	M[15.99]AAQGE[37.95]PGYLAAQSDPGSNSER	0.014199
https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/1896/psm/71789821	470.16356	RPC[57.02]FSALEVDET[470.16]YVPK	0.0072029

Additional Benchmarking of Magnum and MSFragger using Phosphopeptides

Further validation of open search results from Magnum was done by comparing open modification masses returned by Magnum and MSragger with those returned by comet when searching phospho-data from Lawrence et al., 2016 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD003344). These data were acquired from a sample of phosphopeptide-enriched tryptic peptides from a tryptic digest of MCF7 breast cancer cells[23]. We performed closed comet and open Magnum and MSFragger searches on these data and compared the results.

Closed comet searches were done as described in Supplementary Methods, above, using Comet version 2020.01 rev. 0 and with a fixed cysteine carbamidomethylation modification of 57.021464 and a variable phospho-modification of 79.966331 allowed on S, T or Y. Precursor mass tolerance was set to 50 ppm as per the authors original analysis of these data[23]. Isotope error was set to 3. Results were assigned q values using Percolator version 3.05.0.

Open Magnum (version MV1.0.0-alpha5) and MSFragger (version 3.2) searches were done with a fixed cysteine carbamidomethylation modification and other parameters as described in

Supplementary Methods. Search database consisted of the entire reviewed human proteome from Uniprot (https://www.uniprot.org/). Phospho-modifications were intentionally not defined in open searches and thus were only able to be identified as open modifications in the open searches. Figure S12 shows that both Magnum and MSFragger primarily identified peptides containing open modifications of 80 Da from these data. This is expected for a phospho-peptide enriched sample.

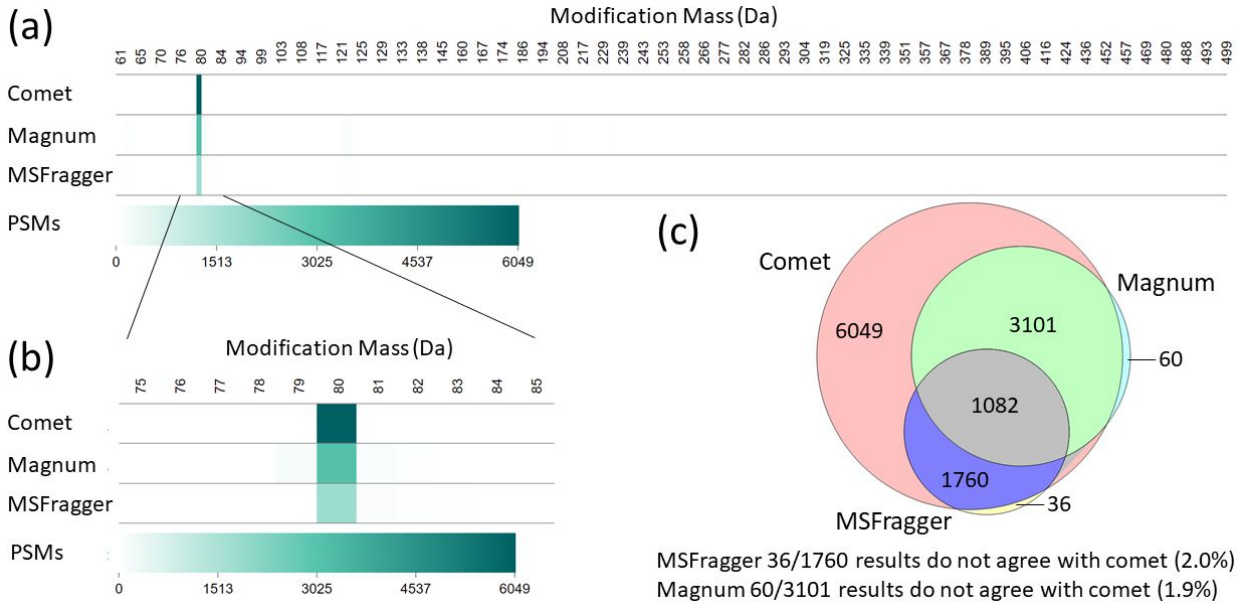


Figure S12: Comparison of phospho-modifications identified by closed comet searching (defined phospho-modification allowed on S, T or Y) versus Magnum and MSFragger open searches (60-500 Da allowed on any amino acid). (a) All modification masses identified by each algorithm. (b) Zoom of (a) showing 75 to 85 Da only. (c) The number and overlap of specific scans resulting in a PSM containing a phospho-modification returned by each algorithm. Results are shown at 1% FDR. Bins are 1 Da wide and phospho-modifications fall within the 80 Da bin. A limelight view of these data can be seen here: https://limelight.yeastrc.org/limelight/go/p94AQVEJCU

A comparison of which specific spectra were found to contain a phospho-modification by each search algorithm is shown in Figure S12c and indicates excellent agreement between the open and closed search methods. Both Magnum and MSFragger results were almost exclusively (~98%) a sub-set of closed search results. Open searching yielded lower sensitivity for this specific modification mass as expected due to the much-increased search space traversed by the open search algorithms. It is also noteworthy that combining results from the two open search engines adds valuable results to the open modification searches in this analysis. As Limelight supports output from multiple search engines users can search their data with multiple algorithms and combine results to increase sensitivity. This technique has previously been shown to add power to closed search approaches[24], however as noted by Lawrence et al.[23] such methods can also lead to aggregation of false positive identifications and care must be taken when applying such approaches.

Supplementary Note 3: Limelight

Description

Limelight is a web application built to analyze, visualize, and share bottom-up MS proteomics data. It is open source and freely available at http://limelight-ms.org.

Central to its design is the separation of data analysis and visualization from the software pipeline that generated the data. Limelight's design makes as few assumptions about the data as possible, providing a generalized platform that fully and equally supports data generated by any MS database search pipeline while providing access to the full stack of proteomics data.

All the native results of each pipeline (e.g., p-values, q-values, Xcorrs, etc.) are available within Limelight, and may be used as filtering and analysis criteria for viewing single searches or combining multiple searches even if those searches used different pipelines. Limelight achieves pipeline-independence via an XML schema, called Limelight XML, that encodes all the results and scores in a generalized way. Limelight XML encodes which programs and versions were run, which scores are present from each program and, importantly, how those scores are to be treated (i.e., are larger or smaller numbers more significant?). Not only can the output of any pipeline be represented, but this provides a level of provenance for the data that enhances reproducibility. Once data are represented as Limelight XML, they may be imported into Limelight via the web interface or via an upload API.

The authors of Limelight have written Limelight XML converters for many popular software pipelines including Comet[4], Percolator[5], the Trans-Proteomic Pipeline (TPP)[25], Crux[26], MSFragger[27], open-pFind[28], Comet-PTM[29], MetaMorpheus[30], MODa[31], TagGraph[32] and Magnum (this paper).

Limelight is written in Java and other standard web technologies, including TypeScript, HTML and CSS. Our GitHub repository resides at: https://github.com/yeastrc/limelight-core and a current list of Limelight XML importers may be seen by visiting https://limelight-ms.readthedocs.io/. The Limelight XML schema may be seen at https://github.com/yeastrc/limelight-import-api/tree/master/xsd.

Experiment Builder

Limelight includes a novel interface to defining the experimental conditions for a set of separate mass spectrometry searches. This is referred to in Limelight as the "experiment builder". Using the experiment builder, users may iteratively build up an experimental design by adding condition groups and then conditions within those condition groups. For example, a user may designate that they have two technical replicates in their experimental design. Then a user may designate they have 6 timepoints in their experimental design. Then a user may add a condition group for "tissue type" and add 3 conditions in this condition group: "heart", "lung", and "liver." (See Figure S13, below). Once an experiment has been structured and searches have been assigned to each cell in the resulting grid, users may view the experiment to automatically compare searches and groups of searches according to their experimental design.

(A)

Experiment Layout

Tech Rep 1	Tech Rep 2
Empty	Empty

(B)

Experiment Layout

	T1	T2	T3	T4	T5	T6
Tech Rep 1	Empty	Empty	Empty	Empty	Empty	Empty
Tech Rep 2	Empty	Empty	Empty	Empty	Empty	Empty

(C)

Experiment Layout

		T1	T2	T3	T4	T5	T6
Heart	Tech Rep 1	Empty	Empty	Empty	Empty	Empty	Empty
	Tech Rep 2	Empty	Empty	Empty	Empty	Empty	Empty
Lung	Tech Rep 1	Empty	Empty	Empty	Empty	Empty	Empty
	Tech Rep 2	Empty	Empty	Empty	Empty	Empty	Empty
Liver	Tech Rep 1	Empty	Empty	Empty	Empty	Empty	Empty
	Tech Rep 2	Empty	Empty	Empty	Empty	Empty	Empty

Figure S13: Iteratively building an experiment using the experiment builder. (A) The user has added two technical replicates. (B) The user has added six time points. (C) The user has added three tissue types. The user may continue to iteratively add condition groups and conditions to a level of infinite complexity to fully represent their experimental design. Once done, the user may click in the empty cells to add searches from the experiment to each cell to organize the data.

Score filters and cutoffs

On all pages (peptide, protein, and modification views), users may filter the data present on the page according to any score present in the respective software pipeline that was used to search the data. For example, E-value, Xcorr, or any other score native to Comet may be used for filtering; and q-value, posterior error probability, or any other score native to Percolator may be used for filtering. If the pipeline was a multistep pipeline, any score from any step may be used.

The currently-used filters are shown at the top of the page, and may be easily changed by clicking on any of the displayed filters to bring up an interface for changing the cutoffs. This interface includes a text box for every type of score present in the native pipeline and entering new values and clicking “Save” will result in those filters being applied.

Single protein view

Wherever protein names are displayed in Limelight, they be clicked to view the single protein view. This view provides data visualizations for a single protein identified in the experiment. This includes the name and description of the protein, its sequence coverage, and a list of all peptides localized to this protein. This list of peptides may be expanded to view all PSMs for each peptide, and each PSM will include native scores and links to view spectra.

A critical aspect of the single protein view is the ability to apply advanced filtering to the peptide list to focus on peptides relevant to a specific question. This peptide list may be filtered by which modifications were identified, whether it was uniquely identified in this protein, peptide sequence, and whether it overlapped specific positions in the protein sequence. The sequence coverage map is interactive, and by clicking positions users may filter for only peptides covering that position (control-click to select multiple positions).

Peptide view

The peptide view provides a peptide-focused view of the experimental results. This view shows all the peptides identified in the experiment, given the current cutoffs (see above). Each listed peptide includes that peptide's sequence, whether it uniquely matched a single protein, in which proteins it was localized, and the spectral count. Each peptide may be clicked on and expanded to reveal all PSMs associated with that peptide, including links to view the annotated spectrum and native scores and annotations from the respective software pipeline.

The peptide list may be filtered according to which modification masses were observed, whether the peptide is unique to a single protein, peptide sequence, or a set of specific proteins (and positions within those proteins). This enables users to perform filtering such as listing only the peptides that contain a phosphorylation and localize to the C-terminal region of all the variants of a given protein identified in the experiment.

Protein view

The protein view provides a protein-focused view of the experimental results. It lists all the proteins identified in the experiment, given the current cutoffs (see above). Each row lists the protein's name, description, sequence coverage, number of peptides, number of unique peptides, and number of PSMs (that meet the current cutoffs). Each row may be clicked to expand to view peptides, and each peptide may be clicked to expand to view PSMs.

Modification view

The modification list view provides a modification-focused view of the experimental results. All modifications identified in the search are displayed in two ways. First the modifications are displayed as a heatmap, where the modification masses are displayed on the x-axis, the currently-shown searches are displayed on the y-axis, and the matrix is shaded according to statistics associated with a given modification mass in a given search. These statistics may be PSM count, scan count, ratio of all PSMs or scans that have that modification mass, or a statistical transformation (described below). The heatmap is interactive, and users may click (and optionally drag) within the visualization to filter which modifications are displayed below in the modification table. The heatmap (and table below) may be further filtered and customized by choosing how to scale the colors, the minimum and maximum modification mass to display, statistical transformations, and filtering based on specific proteins (and positions in those protein to which the modifications must localize).

The modification table below lists each modification on separate rows and may be filtered by interacting with the visualization above. Each row includes the modification mass, links to external modification annotation resources, and the value of the current statistic being displayed in the heatmap for that modification in each of the currently shown searches. Each row may be clicked and expanded to view all proteins, positions in those proteins this given mod localizes, which residues in that protein are modified, and the PSM count for the given modification mass in the respective protein. Each protein may be expanded to view the list of peptides that contain that

modification mass for this protein. Associated with each listed peptide are the N- and C-terminal residues in the protein that flank the peptide, the number of PSMs for that peptide, the positions in the respective protein covered by this modification mass in this peptide, and a list of residues (amino acid codes) modified by this modification mass in this peptide. Each peptide may be expanded to view all PSMs (and associated scores) associated with this peptide, including links to view underlying spectra.

Visualizations and Transformations

The following statistical transformations are available in the modification view data visualization:

- Scaled mean difference: For each mod mass and search display: $(x - \mu) / \mu$, where x is the count or ratio for a mod mass in a search and μ is the mean for that mod mass across all searches.
- Per-mod Z-score: For each mod mass and search display: $(x - \mu) / s$, where x is the count or ratio for a mod mass in a search, μ is the mean for a mod mass across all searches, and s is the standard deviation for this mod mass across all searches.
- Global Z-score: For each mod mass and search display: $(x - \mu) / s$, where x is the count or ratio for a mod mass in a search, μ is the mean for all mod masses across all searches, and s is the standard deviation across all mod masses in all searches.
- Global P-value: For each mod mass and search display: p, where p is the Bonferroni-corrected p-value associated with the global Z-score (the probability of observing a z-score of that magnitude or greater by chance given a normal distribution with the observed mean and standard deviation).
- Global Q-value: For each mod mass and search display: q, where q is the Benjamini-Hochberg q-value associated with the global Z-score (the probability of observing a z-score of that magnitude or greater by chance given a normal distribution with the observed mean and standard deviation).

Two-tailed test of proportions

Users of Limelight may download a report that attempts to identify the most statistically significant modifications in one set of searches versus another. This is done by calculating the ratio of PSMs (or scans) that have the given modification mass and dividing by the total number of PSMs. This ratio is calculated for a given set of searches (e.g., the two biological replicates for a control) and compared against another set of searches (e.g., the two biological replicates for a treatment) and a Z-score is calculated using the canonical test of proportions (see Equation S3). In this equation, x1 is the number of PSMs or scans with the modification mass in the first set of searches, x2 is the number of PSMs or scans with the modification mass in the second set of searches, n1 is the total PSM count for the first set of searches, and n2 is the total PSM count for the second set of searches. This calculation produces a z-score which may be used to compare modification masses for significance and may be converted to a p-value using a lookup table. The p-values in the report are then Bonferroni-corrected to account for multiple hypothesis tests (i.e., the number of modification masses tested).

$$Z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad and \quad p = \frac{x_1 + x_2}{n_1 + n_2}$$

Equation S3: Canonical test of proportions

Supplementary Note 4: Development and validation of adduct discovery pipeline using dicloxacillin, flucloxacillin and HSA

Open modifications unrelated to exposure

Estimates suggest over 50% of spectra remain invisible to traditional “closed” search methods[33]. The prevailing hypothesis is that unidentified spectra constitute peptides not represented in a typical search space: they have undefined post-translational modifications (PTMs), chemical modifications, variant protein sequences or unpredictable cleavage aberrations. Several “open” search strategies have been developed to shed light on these “dark” spectra. These strategies enable mass tolerant database searching and have allowed peptide spectrum matches (PSMs) to be made from a large proportion of previously unassigned spectra in shotgun proteomics data[27–35]. Past open search publications have focused largely on global open modification analyses of large and complex proteomics datasets and show increased numbers of PSMs as well as complex modification landscapes not previously accessible via traditional closed searching.

Limelight is designed to support all bottom-up MS proteomics pipelines and is fully capable of analyzing and visualizing open modification data from unexposed samples. We performed open searching, using 7 open search algorithms, on MS data we acquired from unexposed human serum albumin (HSA) samples (Figure S14).

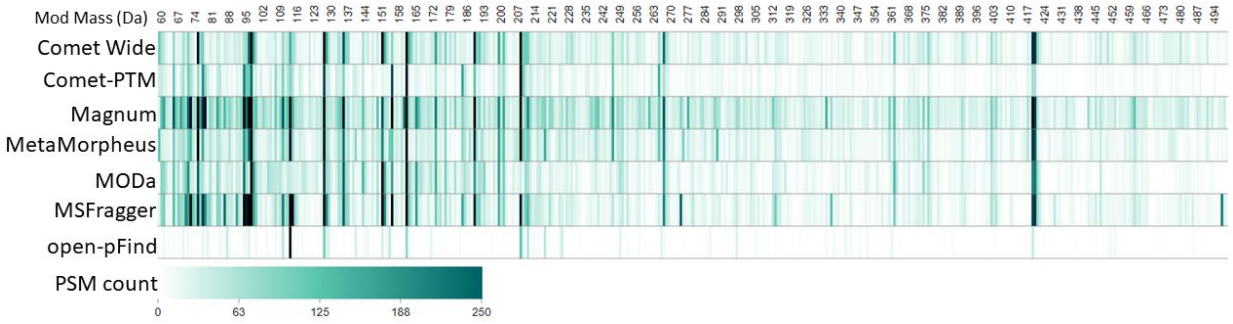


Figure S14: Open modification masses in untreated, purified human serum albumin as identified by 7 different open search algorithms and visualized by Limelight. Open modifications identified in the range of 60 to 500 Da are shown at 1% FDR. A live view of these data is available here: https://limelight.yeastrc.org/limelight/go/Vc6Z8ppwpI

These data show that even in a purified protein sample, open searching results in PSMs containing open modification masses across the entire range of masses available to the search algorithms based on the search parameters they were given.

We then performed both open and closed searches on LC-MS/MS data from 6 untreated and flucloxacillin or dicloxacillin exposed HSA samples. In agreement with previous open search publications, open search analyses resulted in more than twice the number of PSMs as closed searching performed on the same data (Figure S15).

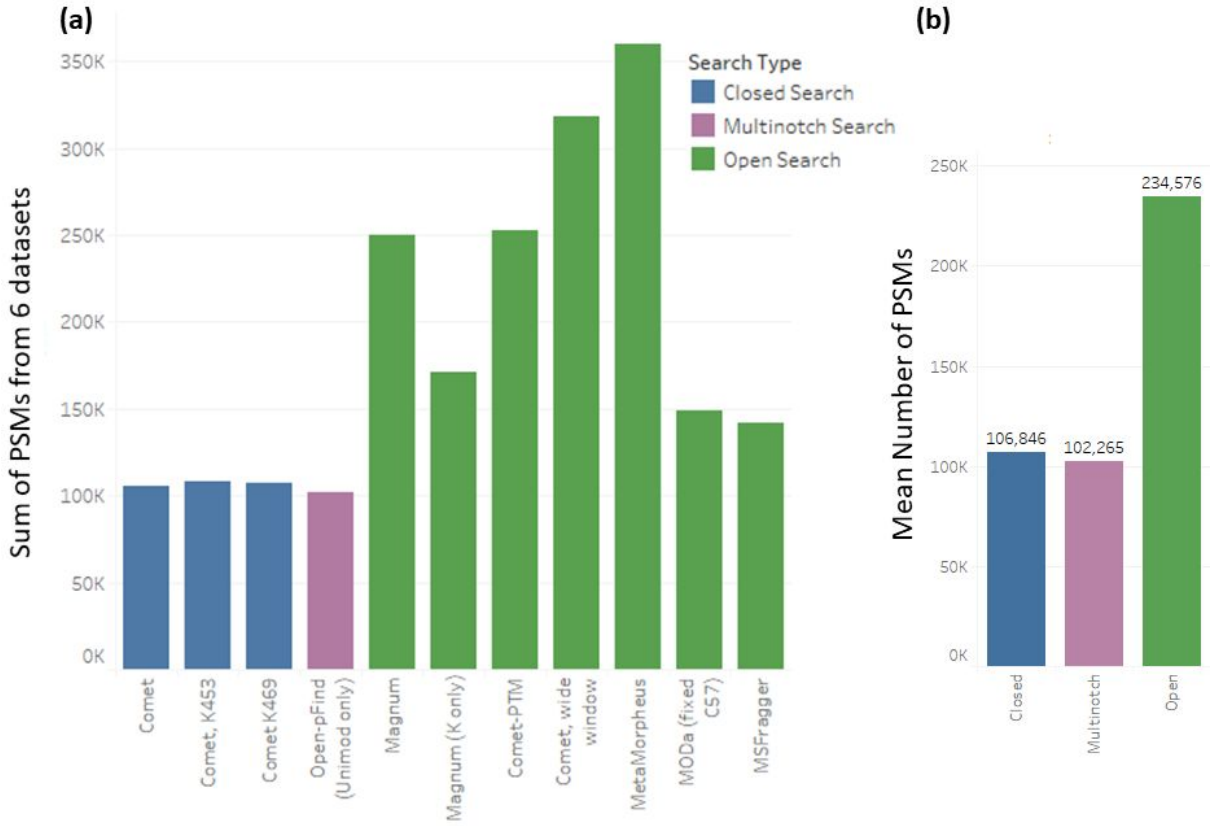


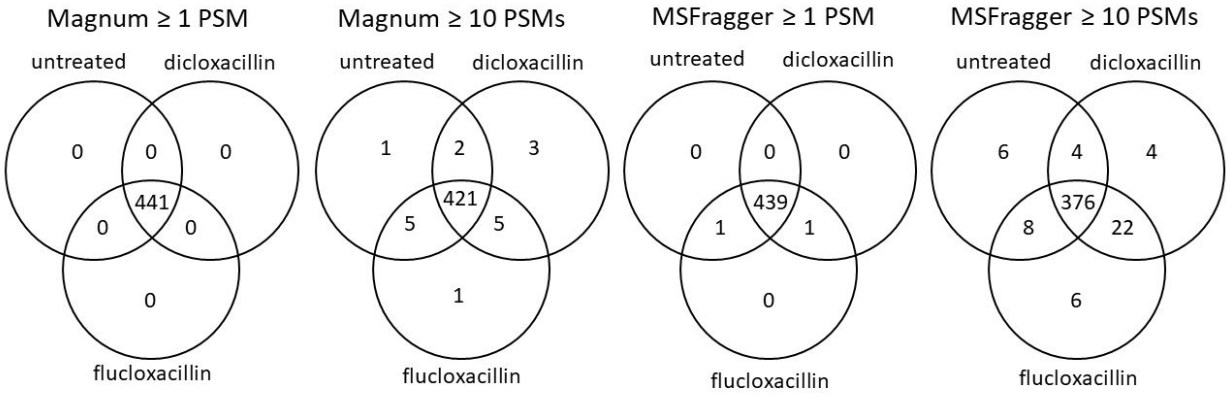
Figure S15: The number of PSMs at 1% FDR resulting from searching 6 untreated and dicloxacillin and flucloxacillin exposed HSA datasets with open and closed search algorithms. (a) The number of PSMs shown is the sum of all 6 datasets for each algorithm. (b) The mean number of PSMs for each algorithm type was calculated from the data in (a). All search engines were configured with 57.02146 on C and 15.99490 on M as variable modifications except MODa which does not allow variable modifications to be defined. Magnum was run allowing open modifications on all residues (Magnum) or restricted to lysines only (Magnum K only). Closed comet searches were performed with the variable modifications previously stated (comet), plus defined variable modifications of the previously published adduct masses of dicloxacillin or flucloxacillin allowed on lysines (comet, K469 and K453).

Discovery of dicloxacillin/flucloxacillin adducts in HSA

Previously published studies[20,21] used multiple methods to determine that dicloxacillin and flucloxacillin produce 469 Da and 453 Da adduct modifications, respectively, on HSA lysine residues (Figure S4).

We acquired untargeted MS data of unexposed, dicloxacillin, and flucloxacillin exposed human serum albumin (HSA) and searched the resulting data using 7 different algorithms. Initially, we compared the open modification masses identified by Magnum in 2 untreated, 2 dicloxacillin treated, and 2 flucloxacillin treated samples and found that the same set of 441 open modification masses (rounding to the nearest integer) were identified in all 6 samples. These masses constituted all the open modification masses available to Magnum based on the search parameters it was given. To reduce this apparent noise, we compared open modification masses identified by ≥10 PSMs in untreated versus treated samples, however none of the adduct masses known to result from dicloxacillin or flucloxacillin exposure were found exclusively in their respective treatment groups. An analysis of the same spectra with an alternative open search tool, MSFragger[27], likewise found at least one PSM for each of the 441 open modification masses

allowed and did not find known exposure related adduct masses uniquely in treated samples using either a 1 or 10 PSM cutoff (Figure S16). These observations are therefore likely inherent to open searching in general and highlight the difficulty in identifying exposure-specific adducts against the background produced by open modification searching.



Search Engine/PSM cutoff	Magnum/ none	Magnum/10 PSMs	MSFragger/ none	MSFragger/ 10PSMs
Num. mod masses common to all treatments	441	421	439	376
Num. mod masses unique to untreated samples	0	1	0	6
Num. mod masses unique to dicloxacillin treated samples	0	3	0	4
Num. mod masses unique to flucloxacillin treated samples	0	1	0	6
Treatment specific mod masses unique to treated samples?	no	no	no	no

Figure S16: Open modification masses identified by Magnum or MSFragger in untreated, dicloxacillin treated or flucloxacillin treated human serum albumin (HSA). The number of open modification masses common or unique to the different treatments is shown. Modification masses known to result from dicloxacillin and flucloxacillin treatment were observed but were not unique to treated samples. Data are filtered at a 1% FDR. The full data presented here can be found at: https://limelight.yeastrc.org/limelight/go/45kI4Zi9cE (Magnum) and https://limelight.yeastrc.org/limelight/go/tTBnwmSdzk (MSFragger).

We overcame these difficulties using the two-tailed test of proportions described in the main manuscript and Supplementary Note 3. We built this method into Limelight and were able to use it to highlight exposure related modifications rapidly and sensitively by comparing untreated with treated samples (or sample groups). These data are presented in the main manuscript for PSMs generated by Magnum. We also searched the same MS data using 6 other open search algorithms for comparison with Magnum. PSMs were imported into Limelight for downstream analysis and a two-tailed test of proportions comparing untreated with dicloxacillin or flucloxacillin treated HSA was performed with PSMs generated by each algorithm.

These analyses resulted in 469 Da (the correct mass) or 471 Da (a monoisotopic mass misassignment of 2 Da) being the most significantly enriched mass for all algorithms tested (Figure S17) except open-pFind which could not identify dicloxacillin or flucloxacillin adducts for the reasons outlined in the main manuscript. Equivalent comparisons with flucloxacillin treated samples resulted in 453 Da being the most significantly enriched mass for all algorithms except open-pFind. Complete data is shown in Supplementary File 2, Sheets 6 and 7 along with links to all data in Limelight.

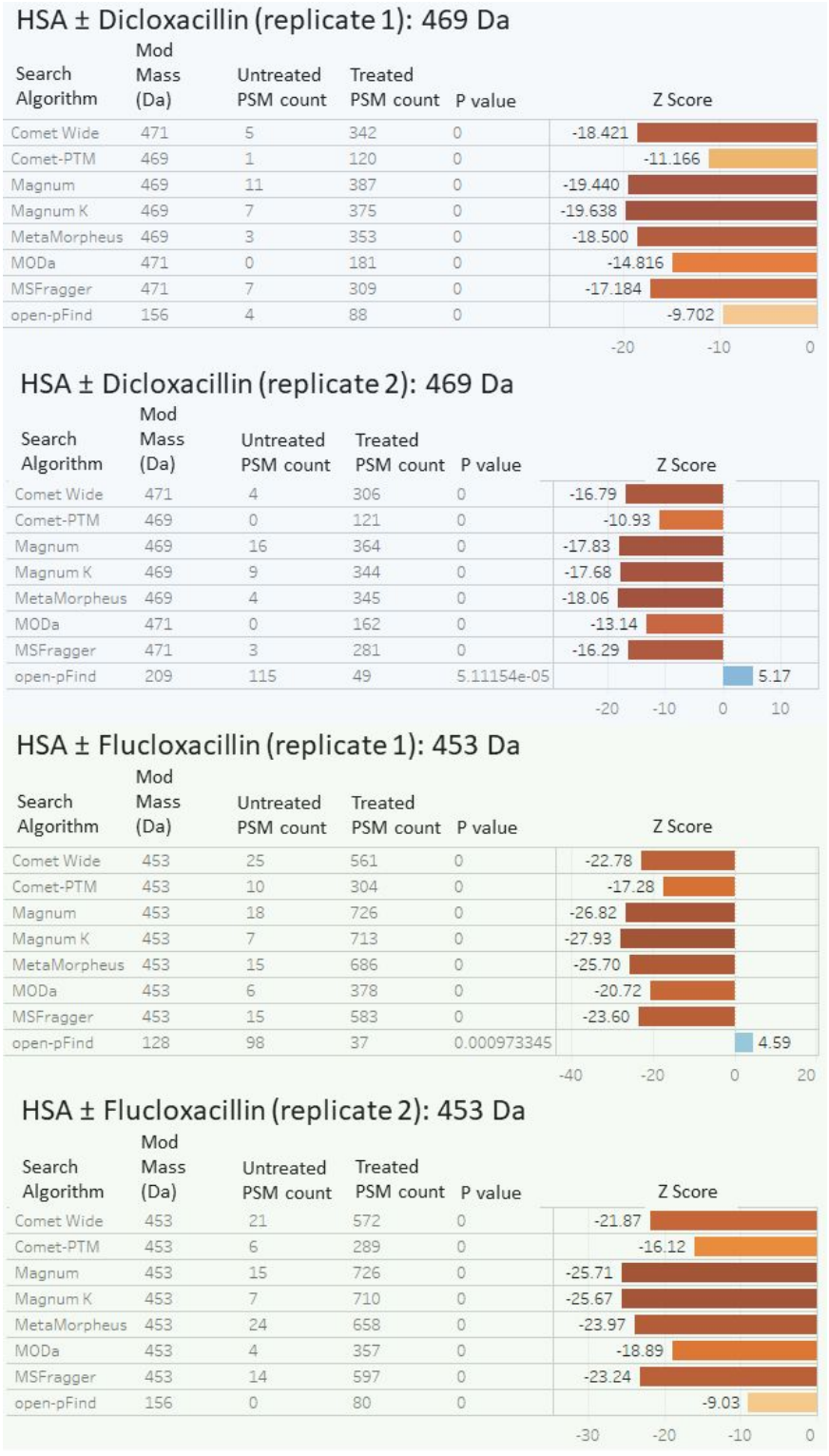


Figure S17: A two-tailed test of proportions performed within Limelight identifies treatment specific adducts in HSA. Tests were done on PSMs identified by 7 different open search algorithms. Results were sorted on the absolute value of the Z score (large to small) followed by the magnitude of the P value (small to large). The top result is shown for each algorithm, representing the most significantly enriched mass found by each algorithm. Magnum was run allowing open modifications on all residues (Magnum) or restricted to lysines only (Magnum K). Data are shown at 1% FDR. Full data is available in Supplementary File 2, Sheets 6 and 7.

In all cases Magnum identified the most treatment related PSMs as well as resulting in the largest Z Score for the correct mass compared to other algorithms. These data show a two-tailed test of proportions comparing treated versus untreated samples, is effective in highlighting exposure specific adducts from PSMs generated by several open search algorithms and that Magnum is the most sensitive algorithm.

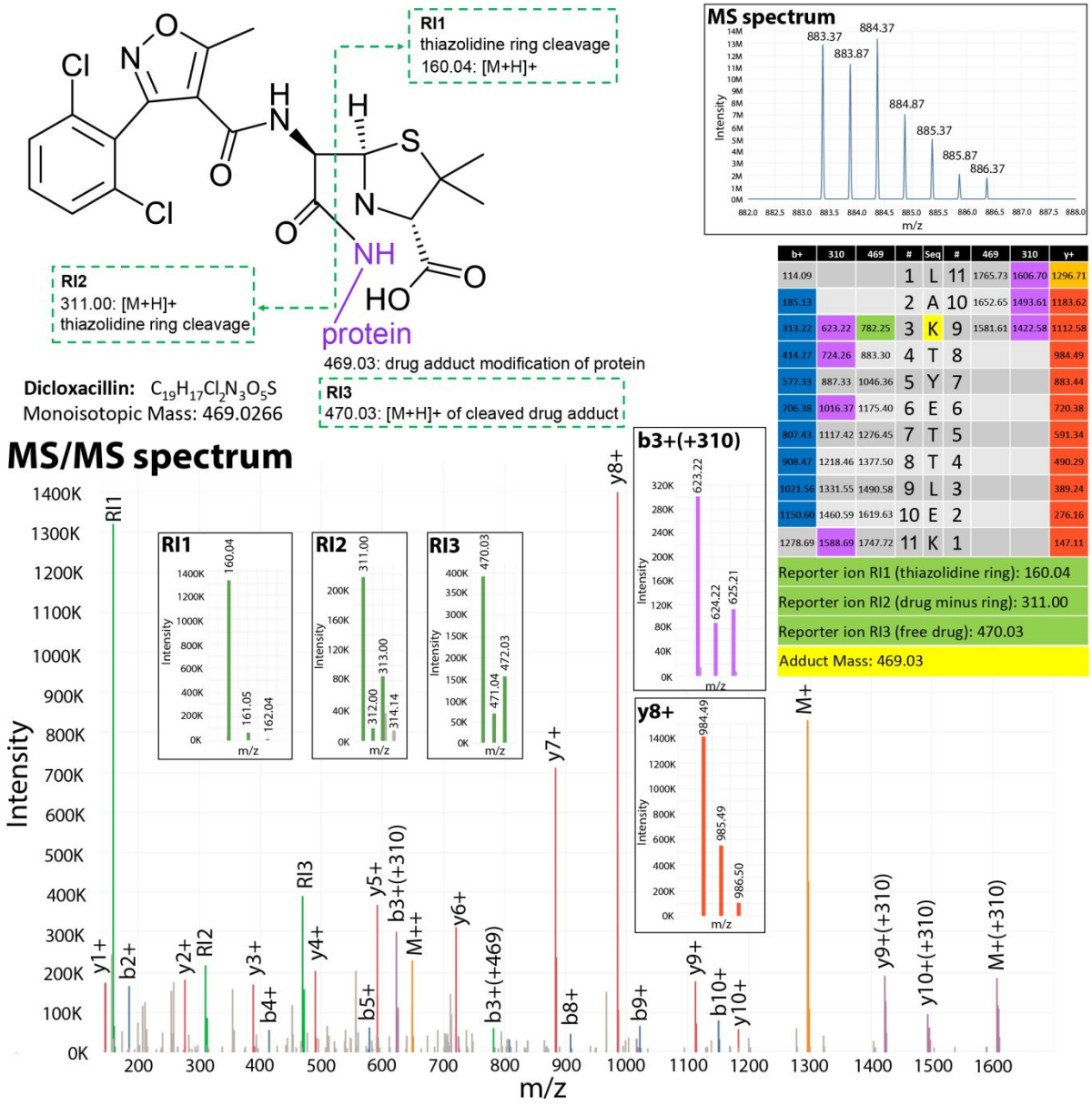
The complete list of +469 Da and +453 Da modified HSA peptides identified by Magnum at 1% FDR is shown in Table S7. All spectra can be manually inspected via Limelight's built in spectrum viewer, Lorikeet[17], using the links in the table caption. A representative, manually annotated spectrum of a dicloxacillin adducted peptide is depicted in the Figure S18 and a similar annotated spectrum of a flucloxacillin adducted peptide is depicted in the Figure S19.

Table S7: Dicloxacillin and flucloxacillin adducted peptides identified by Magnum In purified HSA. Peptides with at least one dicloxacillin (160.04, 311.00 or 470.03) or flucloxacillin (160.04 or 295.03 or 454.06) reporter ion are in black. Peptides with no identified reporter ions are in blue. The 46 dicloxacillin adducted peptides can be viewed on Limelight here: https://limelight.yeastrc.org/limelight/go/avyNUyhojE. The 55 flucloxacillin adducted peptides can be viewed here: https://limelight.yeastrc.org/limelight/go/XFLoAxI0Uw. Equivalent views additionally filtering for presence of reporter ions can be found here https://limelight.yeastrc.org/limelight/go/coQIdEjkRI (for dicloxacillin plus reporter ions) and here https://limelight.yeastrc.org/limelight/go/5aPYwCwMLc (for flucloxacillin plus reporter ions). Data are shown at a percolator q ≤ 0.01.

Dicloxacillin Adducted Peptide Sequence	PSM Count (run 2195)	PSM Count (run 2197)
AFKAWAVAR-(469)	11	12
ASSAKQR-(469)	78	81
ATKEQLK-(469)	97	94
CCAAADPHEC[57]YAK-(469)	0	2
CCAAADPHECYAK-(469)	0	1
D[469]EGKASSAK	1	0
DEGKASSAK-(469)	6	6
EQLKAVMDDFAAFVEK-(469)	1	1
FGERAFK-(469)	10	15
FKDLGEENFK-(469)	0	2
KQTALVELVK-(469)	2	7
KYLYEIAR-(469)	1	0
L[469]VNEVTEFAKTCVADESAENCDK	1	0
LAKTYETTLEK-(469)	30	25
LC[57]TVATLR-(469)	1	1
LDELRDEGK[469]ASSAK	0	1
LDELRDEGKASSAK-(469)	14	8
LKC[57]ASLQK-(469)	9	6
LKCASLQK-(469)	4	5
MPC[57]AEDYLSVVLNQLC[57]VLHEK-(469)	2	0
NECFLQHKDDNPNLPR-(469)	5	4
NLGKVGSK-(469)	61	49

RHPDYSVVLLLR-(469)	0	2
RPC[57]F[469]SALEVDETYVPK	0	1
RPC[57]FS[469]ALEVDETYVPK	6	4
RPC[57]FSAL[469]EVDETYVPK	2	1
RPC[57]FSALEVDETYVPK-(469)	0	1
RPCFSALEVDETYVPK-(469)	0	1
RYKAAFTEC[57]C[57]QAADK-(469)	8	1
RYKAAFTEC[57]CQAADK-(469)	4	1
RYKAAFTECCQAADK-(469)	3	0
SHC[57]IAEVE[469]NDEM[16]PADLPSLAADFVESK	1	0
SHC[57]IAEVENDEMPADLPSL[469]AADFVESK	0	1
SHC[57]IAEVENDEMPADLPSLAADFVESK-(469)	0	2
TC[57]VADESAENC[469]DK	1	2
VGSKC[57]C[57]K-(469)	14	11
VGSKCCK-(469)	2	1
YKAAFTEC[57]C[57]QAADK-(469)	6	10
YKAAFTEC[57]CQAADK-(469)	1	0
YKAAFTECC[57]QAADK-(469)	1	3
YKAAFTECCQAADK-(469)	1	1
Flucloxacillin Adducted Peptide Sequence	**PSM Count (run 2191)**	**PSM Count (run 2193)**
A[453]FKAWAVAR	11	4
AA[453]FTEC[57]C[57]QAADK	1	0
AAFTECC[57]QAADK-(453)	1	4
AAFTECCQAADK-(453)	2	1
AFK[453]AWAVAR	0	1
AFKAWAVAR-(453)	64	67
ASSAKQR-(453)	154	154
ATKEQLK-(453)	107	99
AVMD[453]DFAAFVEK	2	2
AVMDD[453]FAAFVEK	0	1
AVMDDFAAFV[453]EK	0	1
AVMDDFAAFVEK-(453)	2	1
AWAVARLSQR-(453)	1	3
CC[57]AAADPHEC[57]YAK-(453)	0	1
CCAAADPHEC[57]YAK-(453)	2	0
DEGKASSAK-(453)	9	9
EFNAETFTFHADICTLSEK-(453)	0	2
EQLKAVMDDFAAFVEK-(453)	1	0
FGERAFK-(453)	9	19
KQTALVELVK-(453)	24	24

KVPQVSTPTLVEVS[453]RNLGK	0	1
KVPQVSTPTLVEVSR-(453)	1	0
KVPQVSTPTLVEVSRNLG[453]K	2	4
L[453]KECCEKPLLEKSHCIAEVENDEM[16]PADLPSLAADFVESK	1	0
LAKTYETTLEK-(453)	89	80
LDELRDEGK-(453)	0	1
LDELRDEGK[453]ASSAK	0	2
LDELRDEGKASSAK-(453)	71	63
LKC[57]ASLQK-(453)	64	71
LKCASLQK-(453)	1	1
LVRPEV[453]DVMC[57]TAFHDNEETFLK	1	0
MPC[57]AEDYLSVVLNQLC[57]VLHEK-(453)	4	3
MPC[57]AEDYLSVVLNQLCVLHEK-(453)	3	3
NLGKVGSK-(453)	34	40
QNC[57]ELFEQLGEYKFQNALLVR-(453)	0	1
RPC[57]F[453]SALEVDETYVPK	4	2
RPC[57]FS[453]ALEVDETYVPK	6	9
RPC[57]FSA[453]LEVDETYVPK	3	1
RPC[57]FSAL[453]EVDETYVPK	0	1
RPC[57]FSALEVDETYVPK-(453)	7	3
RPCF[453]SALEVDETYVPK	0	1
RPCFSALEVDETYVPK-(453)	1	1
RYK[453]AAFTEC[57]C[57]QAADK	0	1
RYKAAFTEC[57]C[57]QAADK-(453)	0	2
RYKAAFTECC[57]QAADK-(453)	2	1
RYKAAFTECCQAADK-(453)	1	1
SHC[57]IAEVE[453]NDEMPADLPSLAADFVESK	1	0
SHC[57]IAEVENDE[453]MPADLPSLAADFVESK	1	0
VFDEFKPLVE[453]EPQNLIK	2	0
VGSKC[57]C[57]K-(453)	22	24
VGSKCCK-(453)	0	1
VHTEC[57]C[57]HGDLLEC[57]ADDR-(453)	0	2
YKAAFTEC[57]C[57]QAADK-(453)	3	3
YKAAFTEC[57]CQAADK-(453)	3	2
YKAAFTECCQAADK-(453)	4	6



b+	310	469	#	Seq	#	469	310	y+
114.09			1	L	11	1765.73	1606.70	1296.71
185.13			2	A	10	1652.65	1493.61	1183.62
313.22	623.22	782.25	3	K	9	1581.61	1422.58	1112.58
414.27	724.26	883.30	4	T	8			984.49
577.33	887.33	1046.36	5	Y	7			883.44
706.38	1016.37	1175.40	6	E	6			720.38
807.43	1117.42	1276.45	7	T	5			591.34
908.47	1218.46	1377.50	8	T	4			490.29
1021.56	1331.55	1490.58	9	L	3			389.24
1150.60	1460.59	1619.63	10	E	2			276.16
1278.69	1588.69	1747.72	11	K	1			147.11

Figure S18: Annotated MS and MS/MS spectrum of dicloxacillin adducted HSA peptide identified by Magnum. The adduct (structure, top left) is covalently bound to the lysine primary amine (purple). Dicloxacillin is an MS-labile adduct. During peptide fragmentation the adduct itself is cleaved at the thiazolidine ring (green dotted line) releasing reporter ion 1 (RI1) and reporter ion 2 (RI2). The entire adduct can also be cleaved from the peptide releasing reporter ion 3 (RI3). The masses of these ions are independent of the peptide to which the drug is adducted as the ions are derived from adduct fragmentation. For this precursor ion, the MS spectrum indicated an observed mass of 883.37 m/z at charge 2 ($M+2H^{2+}$). The observed precursor mass is thus 1,764.724 Da. The theoretical mass of peptide LAKTYETTLEK is 1,295.697 Da. The adduct modification mass, equal to the mass unexplained by the predicted peptide, is thus 469.03 Da. The peptide sequence and calculated ion series are displayed in the table (center right). The annotated MS/MS spectrum is shown below with inset zoomed panels depicting RI1, RI2, RI3, b3(+310) which is the b3+ ion plus the dicloxacillin adduct minus the thiazolidine ring and is the dominant drug modified ion series, and y8+. Note that the MS spectrum, RI2, RI3 and b3+(+310) all have clear chlorine isotope signatures due to the chlorine in dicloxacillin. R1 has no chlorine signature as the thiazolidine ring has no chlorine. Also note that the dominant b and y ion series are unmodified as the adduct is cleaved off the peptide during the fragmentation step. This prevents correct localization of the adduct by all open search algorithms tested (discussed in Supplementary Note 5). The original identification can be viewed on limelight here: https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2327/psm/139200618

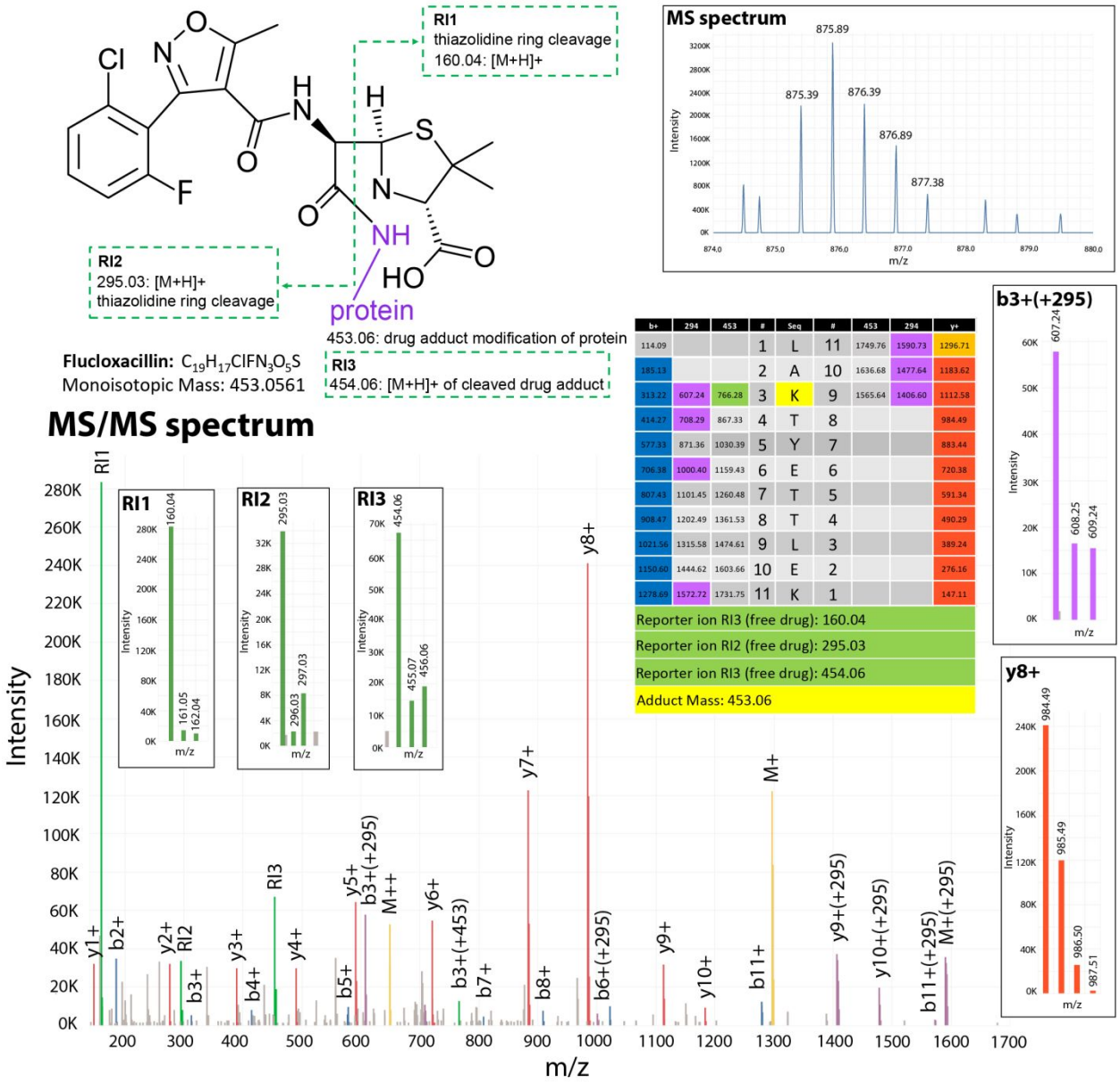


Figure S19: Annotated MS and MS/MS spectrum of flucloxacillin adducted HSA peptide identified by Magnum. The adduct (structure, top left) is covalently bound to the lysine primary amine (purple). Fluoxacillin is an MS-labile adduct. During peptide fragmentation the adduct itself is cleaved at the thiazolidine ring (green dotted line) releasing reporter ion 1 (RI1) and reporter ion 2 (RI2). The entire adduct can also be cleaved from the peptide releasing reporter ion 3 (RI3). The masses of these ions are independent of the peptide to which the drug is adducted as the ions are derived from adduct fragmentation. For this precursor ion, the MS spectrum indicated an observed mass of 875.39 m/z at charge 2 ($M+2H^{2+}$). The observed precursor mass is thus 1,748.76 Da. The theoretical mass of peptide LAKTYETTLEK is 1,295.697 Da. The adduct modification mass, equal to the mass unexplained by the predicted peptide, is thus 453.063 Da. The peptide sequence and calculated ion series are displayed in the table (center right). The annotated MS/MS spectrum is shown below with inset zoomed panels depicting RI1, RI2, RI3, b3(+295) which is the b3+ ion plus the flucloxacillin adduct minus the thiazolidine ring and is the dominant drug modified ion series, and y8+. Note that the R2, RI3 and b3+(+295) all have clear chlorine isotope signatures due to the chlorine in flucloxacillin. R1 has no chlorine signature as the thiazolidine ring has no chlorine. Also note that the dominant b and y ion series are unmodified as the adduct is cleaved off the peptide during the fragmentation step. This prevents correct localization of the adduct by all open search algorithms tested (discussed in Supplementary Note 5). The original identification can be viewed on limelight here: https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2323/psm/138936595

Discovery of dicloxacillin adducts in human plasma

We acquired untargeted MS data of unexposed and dicloxacillin exposed human plasma and searched the resulting data using Magnum. PSMs were imported into Limelight and analyzed as described for HSA above.

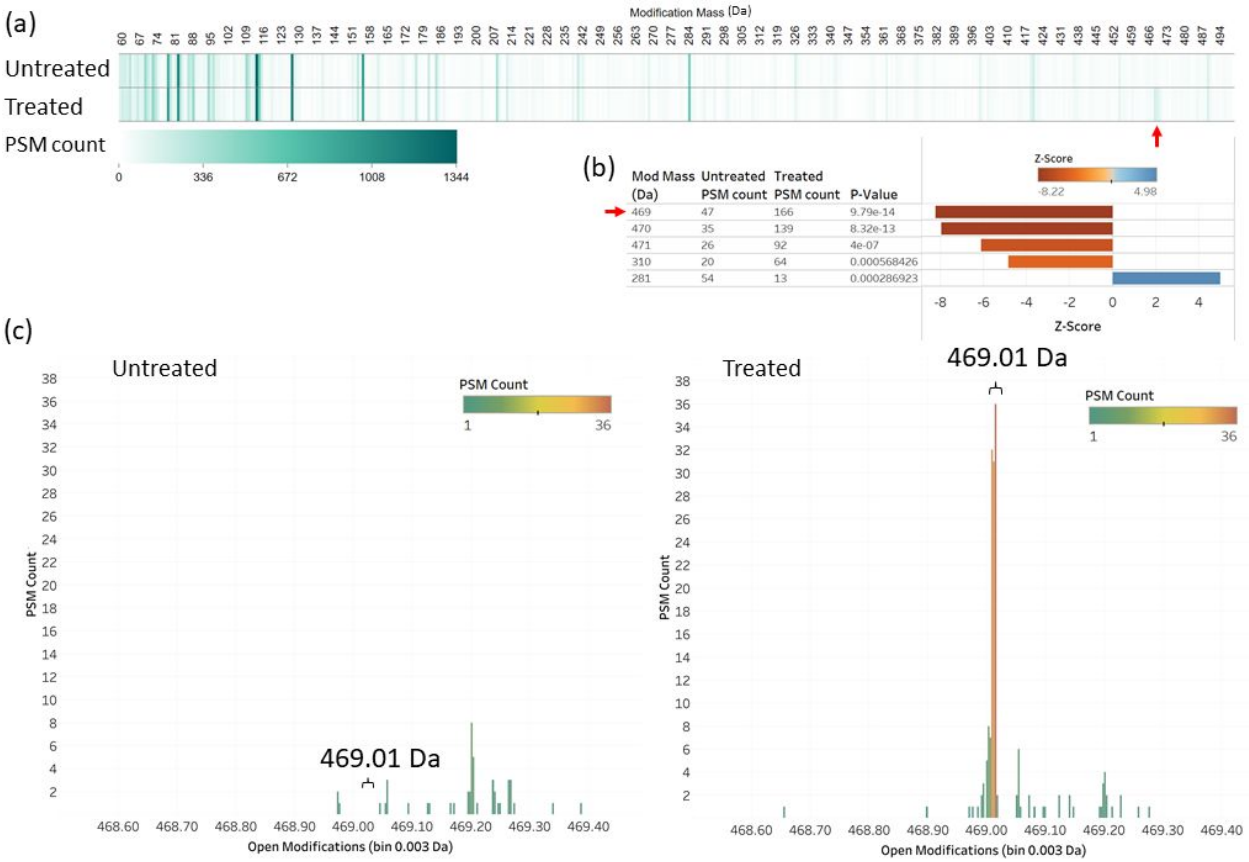


Figure S20: Open modification analysis of untreated and dicloxacillin treated human plasma. PSMs were identified by Magnum and visualized by Limelight. (a) Open modifications identified in the range of 60 to 500 Da are shown. A live view of these data is available here: https://limelight.yeastrc.org/limelight/go/GwAGUaiQKw. (b) A two-tailed test of proportions performed within Limelight identifies 469 Da as the top treatment specific adduct mass in human plasma (red arrow). Results were sorted on the absolute value of the Z score (large to small) followed by the magnitude of the P value (small to large). The top 5 masses are shown. (c) Distribution of observed open-mass identifications within the 469 Da bin. Observed masses in the treated sample peak at 469.01 Da, the exact mass of the dicloxacillin adduct. Observed masses in the unexposed sample do not result in any open-mass identifications in this region (raw data available in Supplementary File 2, Sheets 10 and 11). All results are shown at 1% FDR.

Table S8: Dicloxacillin adducted peptides identified by Magnum in human plasma. Peptides with all 3 dicloxacillin reporter ions (160.04, 311.00 or 470.03) are in black. Remaining peptides are colored blue and have ≥ 1 reporter ion. Data are shown at a percolator q ≤ 0.01 and each peptide was identified by ≥ 2 PSMs. Peptides differing only in adduct location or by leucine/isoleucine substitutions were counted as one. Peptides identified in treated versus untreated samples that contain a 469 Da modification can be viewed on Limelight here: https://limelight.yeastrc.org/limelight/go/UV584MDCHw. *An equivalent view additionally filtering for presence of reporter ions can be found here* https://limelight.yeastrc.org/limelight/go/DyV5xDekzh.

Peptide Sequence	Protein	PSM Count (un-treated)	PSM Count (treated)	Matches treatment specific mass peak (469.01 Da)
ATKEQLK-(469)	ALBU_HUMAN	0	17	yes
ASSAKQR-(469)	ALBU_HUMAN	0	16	yes
FGERAFK-(469)	ALBU_HUMAN	0	14	yes
LDELRDEGKASSAK-(469)	ALBU_HUMAN	0	14	yes
VGSKC[57]C[57]K-(469)	ALBU_HUMAN	0	10	yes
LAKTYETTLEK-(469)	ALBU_HUMAN	0	9	yes
KQTALVELVK-(469)	ALBU_HUMAN	0	8	yes
AWAVARLSQR-(469)	ALBU_HUMAN	0	4	yes
SH[469]C[57]IAEVENDEMPADLP SLAADFVESK	ALBU_HUMAN	0	4	no
YKAAFTEC[57]C[57]QAADK-(469)	ALBU_HUMAN	0	3	yes
C[57]ASIQKFGER-(469)	ALBU_HUMAN	0	2	yes
NLGKVGSK-(469)	ALBU_HUMAN	0	2	yes
C[57]PLMVKVLDAVR-(469)	TTHY_HUMAN	0	2	yes
RLGMFNIQHC[57]K-(469)	A1AT_HUMAN	0	2	no
SKEQLTPLIK-(469)	APOA2_HUMAN	0	2	yes

Supplementary Note 5: Localization of dicloxacillin and flucloxacillin adducts

The main manuscript does not discuss the algorithm-determined localization of dicloxacillin and flucloxacillin HSA adducts. This was done because (1) not all algorithms are able to localize adducts and (2) dicloxacillin and flucloxacillin form labile adducts which typically break off during peptide fragmentation leaving unmodified fragment ions. This is illustrated in Figures S18 and S19. As a result of the adduct being cleaved from the peptide the unmodified ion series is typically more intense (and thus higher scoring) than the adduct modified ion series. As a result of adduct fragmentation there is generally insufficient spectral evidence to localize dicloxacillin and flucloxacillin adducts accurately. Magnum accounts for both labile and stable adducts as described above under Magnum development. In addition, Limelight was designed to simultaneously work with peptides containing both localized and unlocalized open modifications. In the case of labile adducts an unlocalized modification allowing for an unmodified ion series will typically score better than a localized modification within Magnum. For example, 98% of the gold standard PSMs identified by Magnum searches (Supplementary Note 2) scored better as an unlocalized modification even when allowing only unlocalized or lysine localized (the known correct localization) adducts (Table S8). If adducts were allowed on any residue, 91% still scored better as an unlocalized modification within Magnum. Adduct localization performed by MSFragger in combination with PTMProphet[7] was incorrect (not on a lysine) 82% of the time, however 74% of the time the adduct was localized to the first residue of the peptide, which results in an unmodified theoretical y ions series. These results do not reflect the localization capabilities of the algorithms tested but reflect the intrinsic properties of labile adducts, which break off during peptide fragmentation. In addition to being able to handle unlocalized and localized modifications, Limelight incorporates the ability to manually move or set the position of any PTM via the spectral viewer while updating the annotated ions in real time. This allows the user to manually evaluate and define the most likely adduct localization on the peptide if desired.

Table S9: Dicloxacillin and flucloxacillin adduct localization determined by Magnum or MSFragger plus PTMProphet in 2,979 gold standard MS/MS spectra. Correct adduct localization is known to be lysine. Magnum (K only) allowed adducts to be unlocalized or lysine localized. Magnum (all localizations) allows adducts to be unlocalized, or, localized to any amino acid in the matched peptide. MSFragger allows adducts to be localized to any amino acid. Results from comet closed searches allowing adduct localization on lysines only are shown as a positive control. Data are shown at 1% FDR.

Algorithm	% open mods localized to lysine	% open mods localized to N terminal	% open mods unlocalized
Comet closed (K only)	100	4.10	n/a
Magnum (K or unlocalized)	1.58	0.08	98.4
Magnum (all amino acids or unlocalized)	0.21	7.52	91.1
MSFragger (all amino acids)	17.74	73.74	n/a

Supplementary Note 6: CYP3A4/Raloxifene analysis

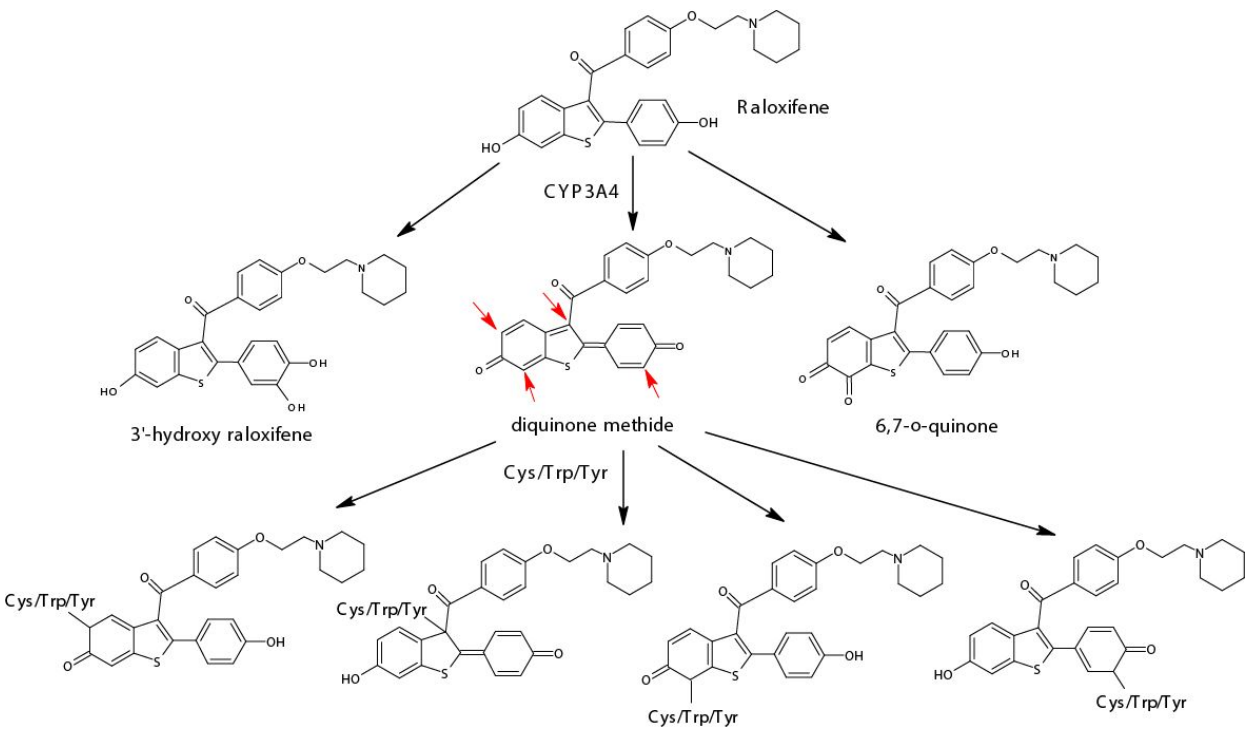


Figure S21: Bioactivation of raloxifene by CYP3A4 and representative adducts formed with cysteine, tryptophan or tyrosine. Red arrows indicate possible sites of diquinone methide which can undergo nucleophilic attack by amino acids to form adducts with a modification mass of 471 Da[37].

Initial Experiments and Skyline Quantification

LC-MS/MS data of raloxifene treated and untreated purified CYP3A4 plus P450-reductase were acquired, searched using Magnum, and analyzed using Limelight. Comparisons of individual raloxifene treated versus untreated replicates, using a two-tailed tests of proportions, resulted in several masses reported as significantly enriched in either untreated or treated samples.

To decrease the significance of masses that varied between replicates while increasing the significance of masses that were constant across replicates, we developed a method able to combine multiple replicates into treatment groups within Limelight (see Experiment Builder, Supplementary Note 3). A two-tailed test of proportions analyzing Magnum data on untreated versus raloxifene treated groups resulted in 471 Da being the most significantly enriched mass in the raloxifene treated sample group (Table S10).

Table S10: A two-tailed test of proportions identifies a 471 Da raloxifene specific adduct mass in CYP3A4 and P450-reductase. PSMs were generated using Magnum and analyzed using Limelight. Results were sorted on the absolute value of the Z score (large to small) followed by the magnitude of the P value (small to large). The top 5 masses are shown, representing the most significantly enriched masses found in either the treated sample groups (negative Z scores) or the untreated sample group (positive Z scores). Magnum was run allowing for open masses on any amino acid. Data are shown at 1% FDR. Full data is available in Supplementary File 2 Sheets 8 and 9 and a live view the data is available at: https://limelight.yeastrc.org/limelight/go/wwa741ZTTE (click [View Replicate ZScore Report] to view table).

mod mass	untreated PSM count	treated PSM count	z-score	p-value
471	**14**	**146**	**-10.485762**	**0**
293	367	183	7.78277453	3.13349E-12
298	594	375	6.95595554	1.54427E-09
72	490	306	6.44581951	5.07051E-08
115	174	79	5.92416284	1.38427E-06

We searched the same MS data using 6 other open search algorithms. PSMs were imported into Limelight and a two-tailed test of proportions comparing untreated versus raloxifene treated sample groups was performed with PSMs generated by each algorithm. Of the other algorithms used, MetaMorpheus also found 471 Da as being the most significantly enriched mass in the raloxifene treated sample group (Figure S22a). However, Magnum identified twice as many PSMs with that modification mass in treated samples than MetaMorpheus (146 versus 72) and fewer 471 Da masses in untreated samples (14 versus 16). The treated:untreated ratio of 471 Da PSMs for MetaMorpheus was 5:1 indicating about 80% of the 72 PSMs with a 471 Da modification in treated samples were treatment specific. In contrast an equivalent Magnum search had a 10:1 ratio as described in the main manuscript. Other algorithms identified a range of masses as being significantly enriched in treated or untreated samples, however there were typically many PSMs for each of these masses in both treated and untreated samples and the resulting data did not indicate clear treatment specific adduct masses. For example, Magnum and MSFragger reported a similar number of PSMs containing 471 Da modification masses in treated samples, however 471 Da modification masses were also common in PSMs identified by MSFragger in untreated samples leading to a treated to untreated ratio of 3:1. This suggests that 1 in 3 PSMs containing a 471 Da modification mass identified by MSFragger were likely unrelated to exposure. For this experiment, therefore, the advantages of Magnum, which was designed specifically for xenobiotic protein-adduct detection, were required to detect and distinguish raloxifene adducts in a background of other masses. A complete summary of results from all algorithms and links to all data in Limelight can be found in Supplementary File 2, Sheets 8 and 9.

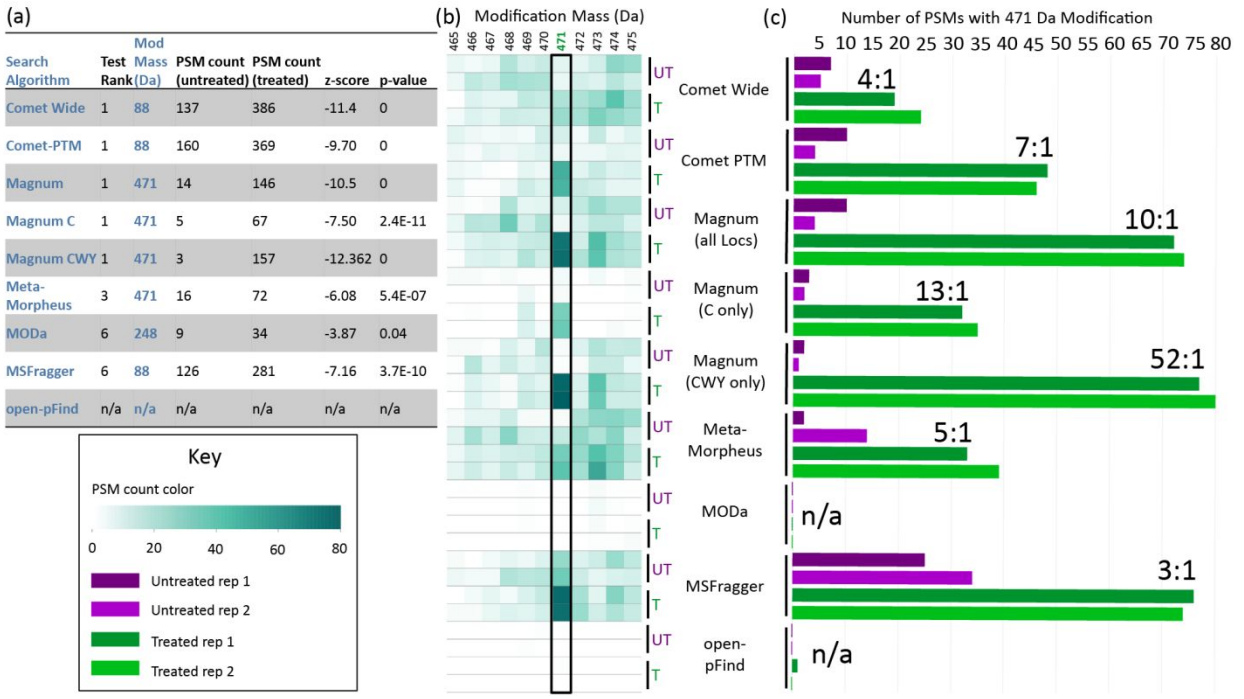


Figure S22: Identification of novel raloxifene adducts in CYP3A4 and P450-reductase. (a) A two-tailed test of proportions performed within Limelight on treatment groups from PSMs identified by 7 open search algorithms. Results were sorted on the absolute value of the Z score (large to small) followed by the magnitude of the P value (small to large). The top significant (p≤0.05) negative Z score is shown for each algorithm, representing the most significantly enriched mass found by each algorithm in the raloxifene treated samples. Magnum was run allowing for open masses on any amino acid (Magnum), restricted to cysteines only (Magnum C), or restricted to cysteine, tryptophan, and tyrosine (Magnum CWY). (b) Heatmap of the number of PSMs identified by 7 open search algorithms containing a 465 through 475 Da open modification masses. For each algorithm counts are shown for 2 untreated replicates and 2 treated replicates. (c) Number of PSMs identified by 7 open search algorithms containing a 471 Da open modification mass. Ratios to the right of the bars represent untreated:treated PSMs containing a 471 Da open modification mass. Data are shown at 1% FDR. Full data is available in Supplementary File 2, Sheets 8 and 9 and a live view of (b) is available at: https://limelight.yeastrc.org/limelight/go/N8sMlKwtQm.

All 146 PSMs containing a 471 Da mass returned by a fully unrestricted (modifications allowed on any amino acid) Magnum search were manually evaluated (Supplementary File 2, Sheet 2) and it was observed that close to 90% of 471 Da adducts were on cysteine, tryptophan or tyrosine. Further Magnum searches were thus performed restricting open modifications to cysteines (C), tryptophan (W), and tyrosine (Y).

Table S11 shows the initial set of residues identified by Magnum-CWY searches as having a 471 Da modification in ≥ 2 PSMs. The number of PSMs at each location is recorded as well as the ion signal quantified in in Skyline (see below). Full data plus Limelight links can be found in Supplementary File 2, Sheet 3.

Table S11: All locations identified as modified by 471 Da adduct masses by Magnum-CWY in CYP3A4 and P450-reductase in initial experiments by ≥ 2 PSMs. PSM counting and Skyline quantification of extracted ion chromatograms were performed to estimate the degree of modification at each location. Full data is available in Supplementary File 2, Sheet 2) and data is plotted in Figures S21, S22 and S23 below. Data is shown at a Percolator calculated PSM level q≤0.01.

Protein Name	PDB numbering modification location	Fasta numbering modification location	Total PSMs at that location	Skyline Quantified Ion Signal at that location
CYP3A4	Y53	Y43	2	0.0000111848
CYP3A4	C58	C48	7	0.000408657
CYP3A4	Y75	Y65	3	0.0000385597
CYP3A4	C98	C88	20	0.001456899
CYP3A4	W126	W116	65	0.003233199
CYP3A4	Y152	Y142	8	0.000158244
CYP3A4	Y430	Y420	2	0.0000243249
CYP3A4	Y432	Y422	8	0.000229608
CYP3A4	C468	C458	19	0.001078295
Reductase	C472	C472	8	0.000118627
Reductase	C630	C630	4	0.000186243

Data were quantified by spectral counting and additionally, extracted ion chromatograms (XICs) were produced and quantified in Skyline[8,9,36]. Full Skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url. Skyline XICs showed treatment specific signal for all peptides identified by Magnum as exclusively in treated samples. Only one peptide, VWGFYDGQQPVLAITDPDM[+16]IKTVLVKEC[+471.2]YSVFTNR, was found to be unrelated to raloxifene treatment. This peptide, identified in a single untreated PSM, was found by both PSM counting and Skyline quantification to be unrelated to raloxifene treatment (Figure S22) and was included as the single non-raloxifene specific peptide had the same 471 Da adduct location (C98) as multiple treatment-specific identifications in separate peptides. All other peptides showed treatment specific 471 Da modifications.

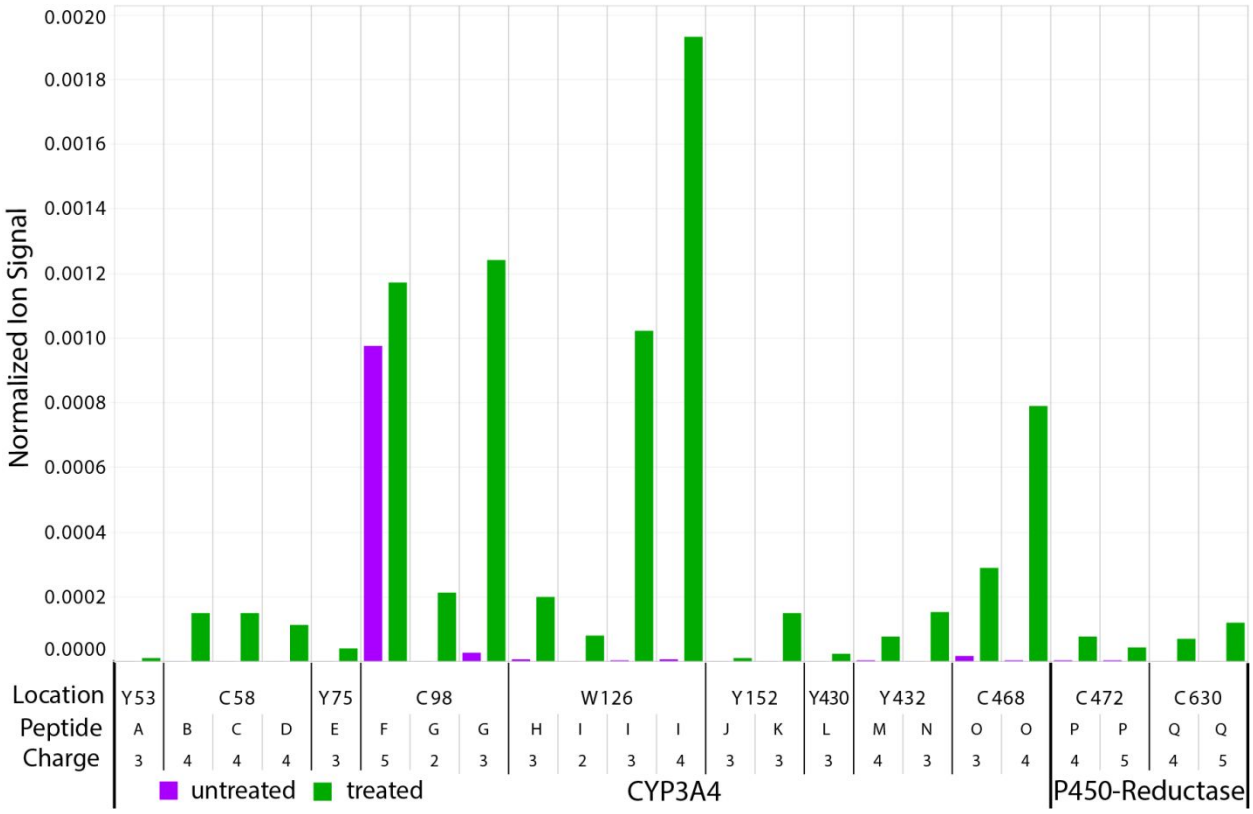


Figure S23: Total normalized ion signal in treated and untreated samples, quantified in Skyline, for individual precursor ions corresponding to peptides identified in CYP3A4 and P450-reductase at each location found by Magnum-CWY searches as modified by a 471 Da adduct mass in initial experiments by ≥ 2 PSMs. Full data are available in Supplementary File 2, Sheet 4. Full Skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url. Note that peptide F (VWGFYDGQQPVLAITDPDM[+16]IKTVLVKEC[471]YSVFTNR) is not expected to be treatment specific and is only included as the residue (C98) predicted to be modified by 471 Da in one untreated PSM was also identified as modified by 471 Da in 20 other treatment specific PSMs in a different peptide (peptide G: EC[471]YSVFTNR) covering the same residue. The modifications in the other peptide are clearly treatment specific. Table S12 shows the full peptide sequence corresponding to the peptide letter abbreviations used in this figure.

Table S12: Full peptide sequences corresponding to the peptide letter abbreviations used in Figure S23.

Protein	Peptide Modified Sequence	Peptide	Precursor Charge	PDB numbering modification location	Fasta numbering modification location
CYP3A4	LGIPGPTPLPFLGNILSY[+471.2]HK	A	3	53	43
CYP3A4	GFC[+471.2]M[+16]FDMEC[+57]HKK	B	4	58	48
CYP3A4	GFC[+471.2]MFDM[+16]EC[+57]HKK	C	4	58	48
CYP3A4	GFC[+471.2]M[+16]FDM[+16]EC[+57]HKK	D	4	58	48
CYP3A4	VWGFY[+471.2]DGQQPVLAITDPDM[+16]IK	E	3	75	65
CYP3A4	VWGFYDGQQPVLAITDPDM[+16]IKTVLVKEC[+471.2]YSVFTNR	F	5	98	88
CYP3A4	EC[+471.2]YSVFTNR	G	2	98	88
CYP3A4	EC[+471.2]YSVFTNR	G	3	98	88
CYP3A4	SAISIAEDEEW[+471.2]K	H	3	126	116
CYP3A4	SAISIAEDEEW[+471.2]KR	I	2	126	116
CYP3A4	SAISIAEDEEW[+471.2]KR	I	3	126	116
CYP3A4	SAISIAEDEEW[+471.2]KR	I	4	126	116
CYP3A4	EMVPIIAQY[+471.2]GDVLVR	J	3	152	142
CYP3A4	EM[+16]VPIIAQY[+471.2]GDVLVR	K	3	152	142
CYP3A4	DNIDPY[+471.2]IYTPFGSGPR	L	3	430	420
CYP3A4	NKDNIDPYIY[+471.2]TPFGSGPR	M	4	432	422
CYP3A4	DNIDPYIY[+471.2]TPFGSGPR	N	3	432	422
CYP3A4	VLQNFSFKPC[+471.2]K	O	3	468	458
CYP3A4	VLQNFSFKPC[+471.2]K	O	4	468	458
P450-Reductase	VHPNSVHIC[+471.2]AVAVEYEAK	P	4	472	472
P450-Reductase	VHPNSVHIC[+471.2]AVAVEYEAK	P	5	472	472
P450-Reductase	LIHEGGAHIYVC[+471.2]GDAR	Q	4	630	630
P450-Reductase	LIHEGGAHIYVC[+471.2]GDAR	Q	5	630	630

Quantification of 471 Da modifications by spectral counting (Figure S24) and by Skyline quantification of XICs (Figure S25) produced similar results and thus spectral counting was performed when combining data from all raloxifene adduct experiments.

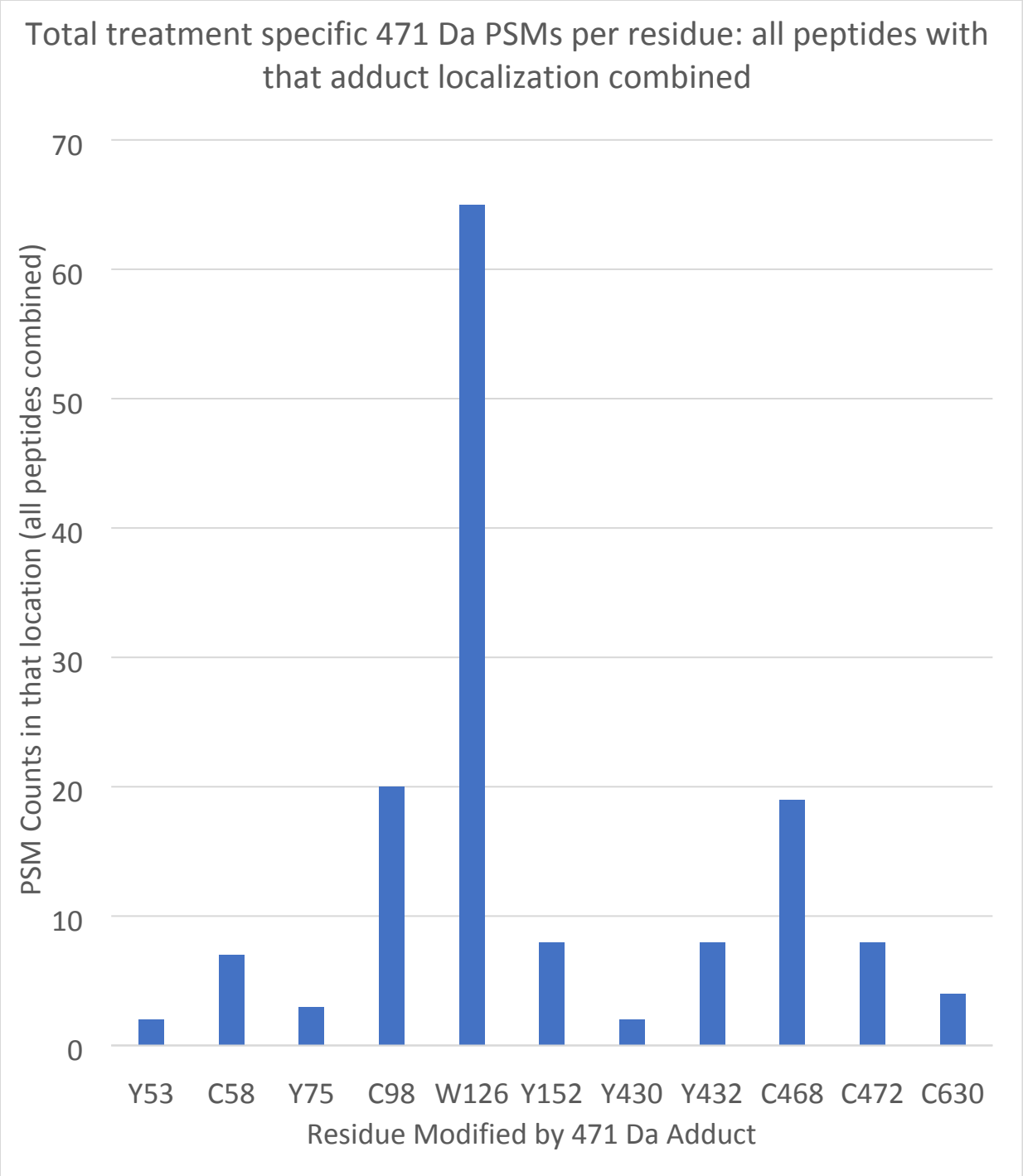


Figure S24: Total PSMs identified in CYP3A4 and P450-reductase at each location found by Magnum-CWY searches as modified by a 471 Da adduct mass in initial experiments by ≥ 2 PSMs. Data is shown at a Percolator calculated PSM level q≤0.01. Full data is available in Supplementary File 2, Sheet 3.

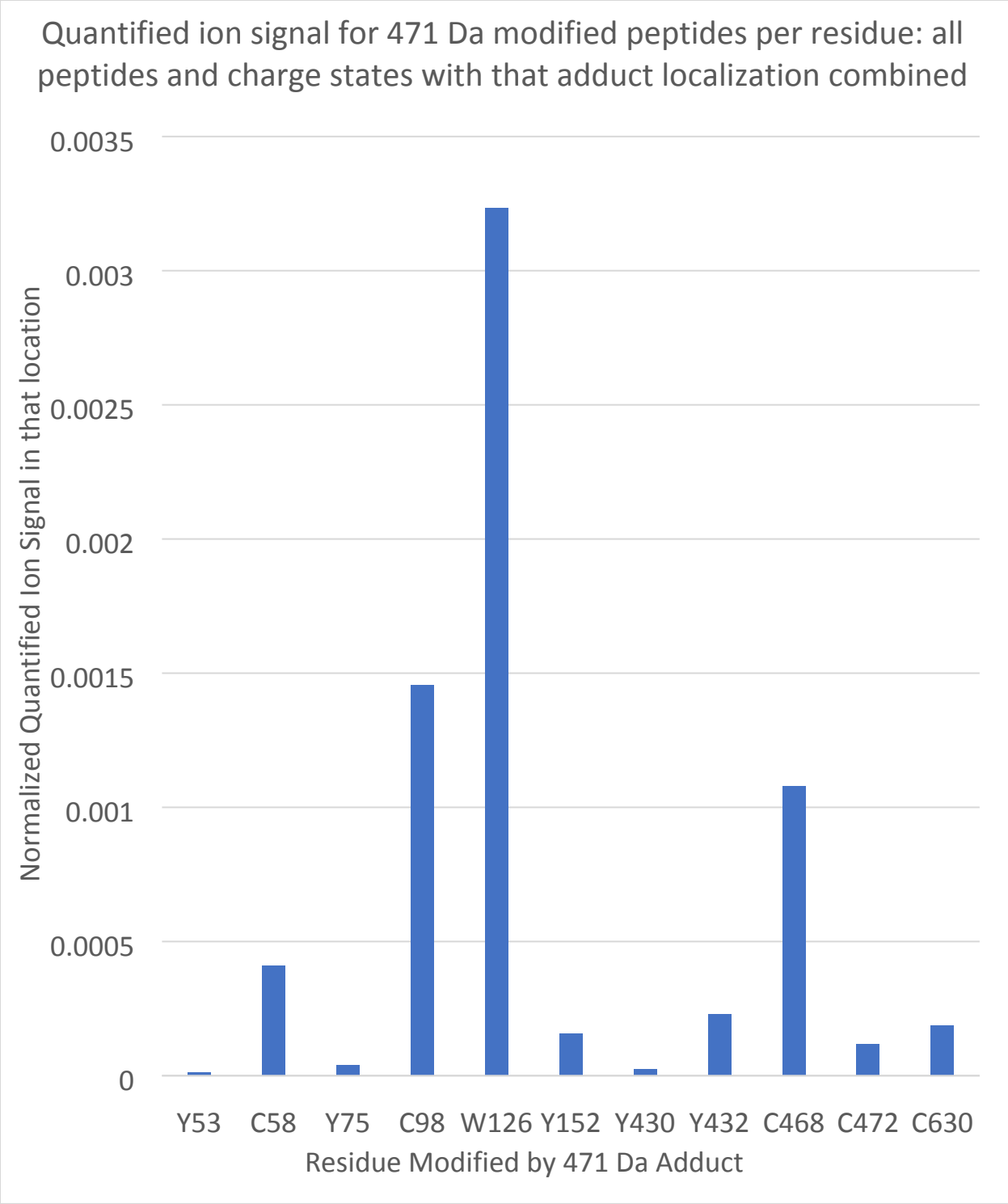


Figure S25: Total normalized ion signal quantified in Skyline for all locations identified in CYP3A4 and P450-reductase at each location found by Magnum-CWY searches as modified by a 471 Da adduct mass in initial experiments by ≥ 2 PSMs. Full data is available in Supplementary File 2, Sheet 4. Full Skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url. The final results, combining all raloxifene replicates, are presented in Figure S30 and in tabular form below (Table S13). The complete analysis is available in Supplementary File 2, Sheets 3-5.

Several raloxifene adducts result in multiple distinct chromatographic peaks

During quantification of raloxifene specific precursor ions, we observed that several raloxifene (471 Da) adducted peptides were associated with distinct chromatographic peaks. The three most obvious examples were W126, C98 and C468, which were also the ions with the largest signals (Figures S26-S28). The following CYP3A4 peptides were observed to produce >1 chromatographic peak: K.ECYSVFTNR.R (3 peaks); K.SAISIAEDEEWKR.L (4 peaks); K.NKDNIDPYIYTPFGSGPR.N (2 peaks); K.DNIDPYIYTPFGSGPR.N (2 peaks); K.DNIDPYIYTPFGSGPR.N (2 peaks); R.VLQNFSFKPCK.E (2 peaks). P450-reductase peptide K.LIHEGGAHIYCCGDAR.N also produced 2 peaks. All other peptides (e.g. Figure S29) produced 1 chromatographic peak under the experimental conditions used.

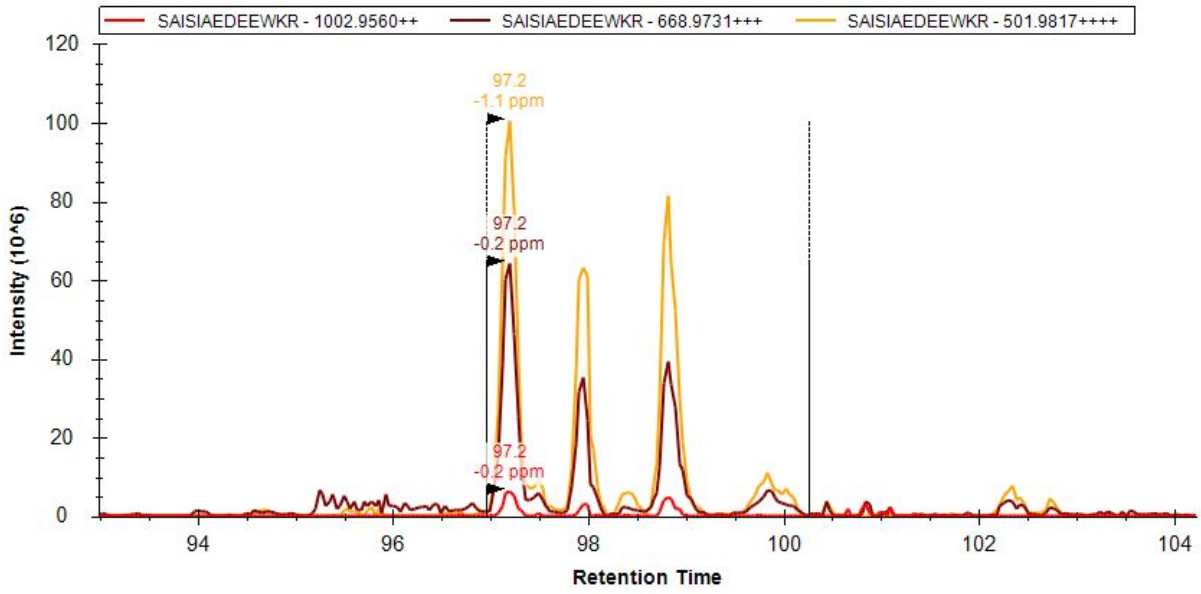


Figure S26: Precursor ions resulting corresponding to the W126 peptide (SAISIAEDEEW[471]KR) elute as 4 distinct chromatographic peaks. Skyline derived extracted ion chromatograms are shown of 2+, 3+ and 4+ precursor ions. Full skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url

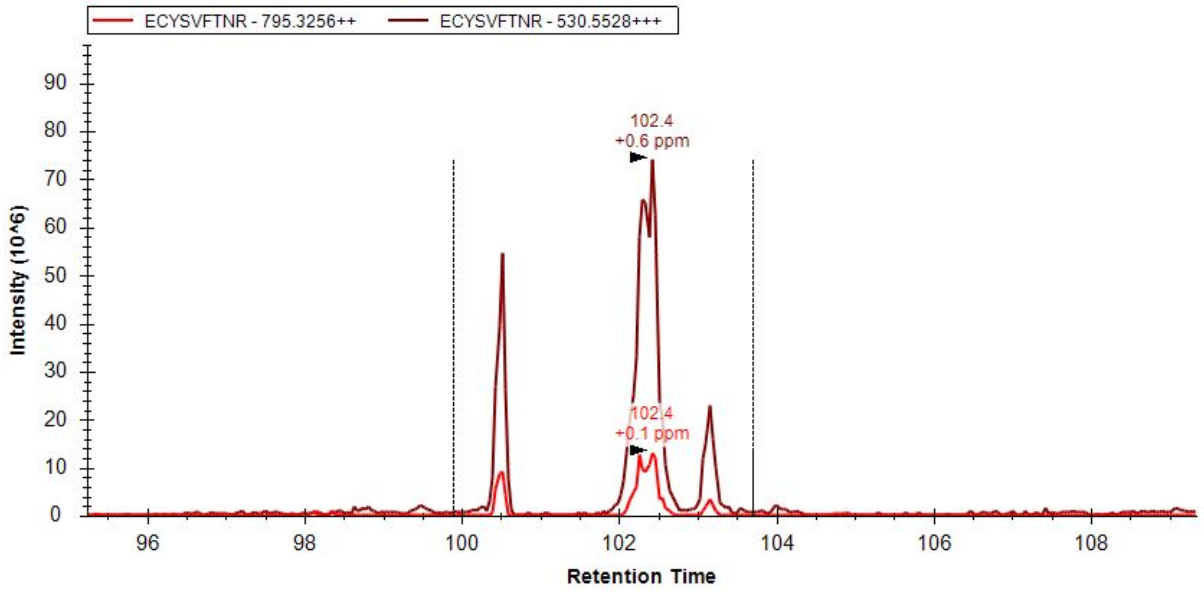


Figure S27: Precursor ions resulting corresponding to the C98 peptide (EC[471]YSVFTNR) elute as 3 distinct chromatographic peaks. Skyline derived extracted ion chromatograms are shown of 2+ and 3+ precursor ions. Full skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url

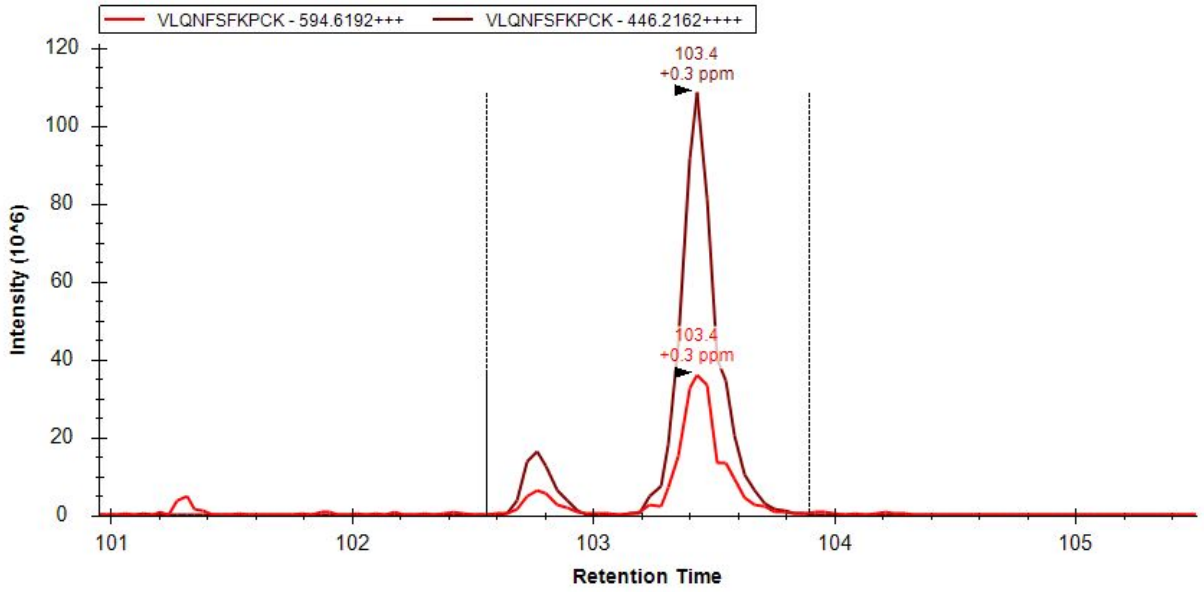


Figure S28: Precursor ions resulting corresponding to the C468 peptide (VLQNFSFKPC[471]K) elute as 2 distinct chromatographic peaks. Skyline derived extracted ion chromatograms are shown of 3+ and 4+ precursor ions. Full skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url

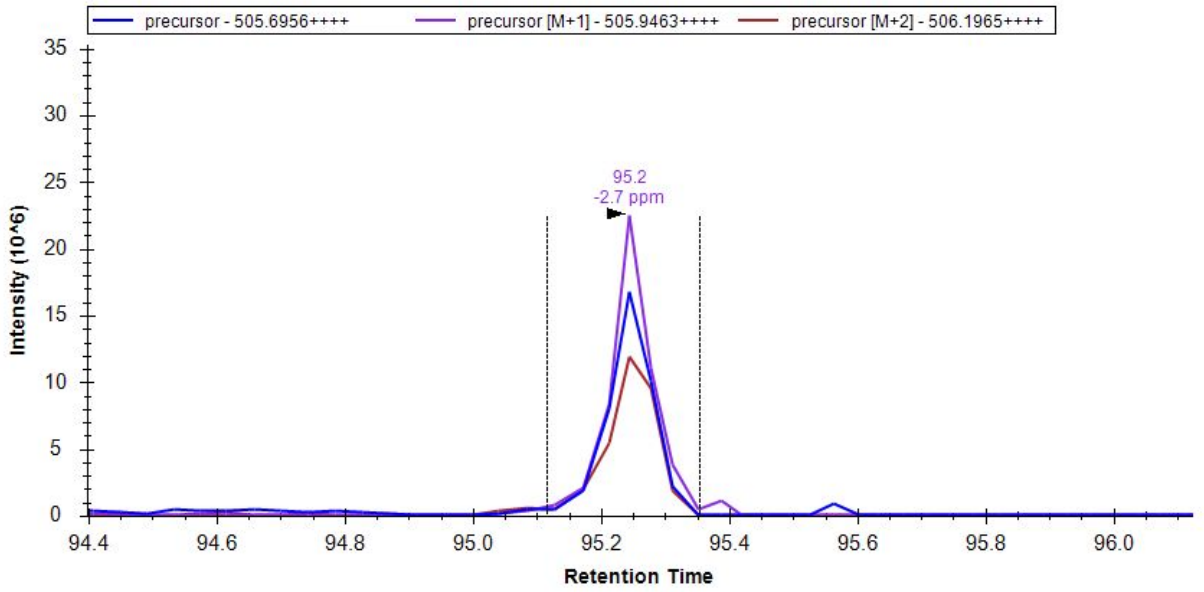


Figure S29: Precursor ions resulting corresponding to the C58 peptide (GFC[471]M[16]FDMEC[57]HKK) elute as 1 single distinct chromatographic peak. Skyline derived extracted ion chromatograms are shown of the 4+ precursor ion. Full skyline sessions are available on Panorama here: https://panoramaweb.org/CYP3A4-raloxifene.url

Previously published work[37] identified 4 different positions in the raloxifene metabolite, diquinone methide, subject to nucleophilic attack and the presence of multiple distinct chromatographic peaks could be explained by adducts on these differing positions (Figure S21).

Further raloxifene experiments

To increase peptide the depth of our CYP3A4/raloxifene analysis we performed two additional sets of MS analyses on our raloxifene treated and untreated purified CYP3A4 plus P450-reductase samples. Firstly, a longer tryptic digest was performed (see “extra digest” in Materials and Methods) and secondly a higher acetonitrile LC gradient was used for select CYP3A4 samples (see “highB” in Materials and Methods). Results from all CYP3A4 plus P450-reductase experiments combined are presented below.

Table S13: Magnum identifies multiple 471 Da protein adducts in CYP3A4 and P450-reductase after exposure to raloxifene. Abundance of all modifications is shown relative to W126, which was observed in 121 PSMs. All results shown have a Percolator calculated PSM level q≤0.01 and were identified by ≥3 PSMs. Raw and data and visualizations are available on Limelight at: https://limelight.yeastrc.org/limelight/go/0UjwIJNz45 (CYP3A4) and https://limelight.yeastrc.org/limelight/go/ypAgoCB3IE (reductase).

Protein	PDB numbering modification location	Fasta numbering modification location (used in Limelight)	Total treatment specific PSMs in that location	Relative Abundance
CYP3A4	**Y53**	43	18	0.14876
CYP3A4	**C58**	48	8	0.06612
CYP3A4	**Y75**	65	9	0.07438
CYP3A4	**C98**	88	30	0.24793
CYP3A4	**W126**	116	121	1
CYP3A4	**Y152**	142	76	0.6281
CYP3A4	**Y399**	389	40	0.33058
CYP3A4	**Y407**	397	23	0.19008
CYP3A4	**Y430**	420	21	0.17355
CYP3A4	**Y432**	422	25	0.20661
CYP3A4	**C442**	432	3	0.02479
CYP3A4	**C468**	458	49	0.40496
Reductase	**Y84**	84	11	0.09091
Reductase	**Y178**	178	7	0.05785
Reductase	**Y259**	259	4	0.03306
Reductase	**Y269**	269	13	0.10744
Reductase	**Y373**	373	13	0.10744
Reductase	**Y374**	374	9	0.07438
Reductase	**Y387**	387	5	0.04132
Reductase	**C472**	472	8	0.06612
Reductase	**Y564**	564	3	0.02479
Reductase	**C630**	630	3	0.02479
Reductase	**Y672**	672	8	0.06612

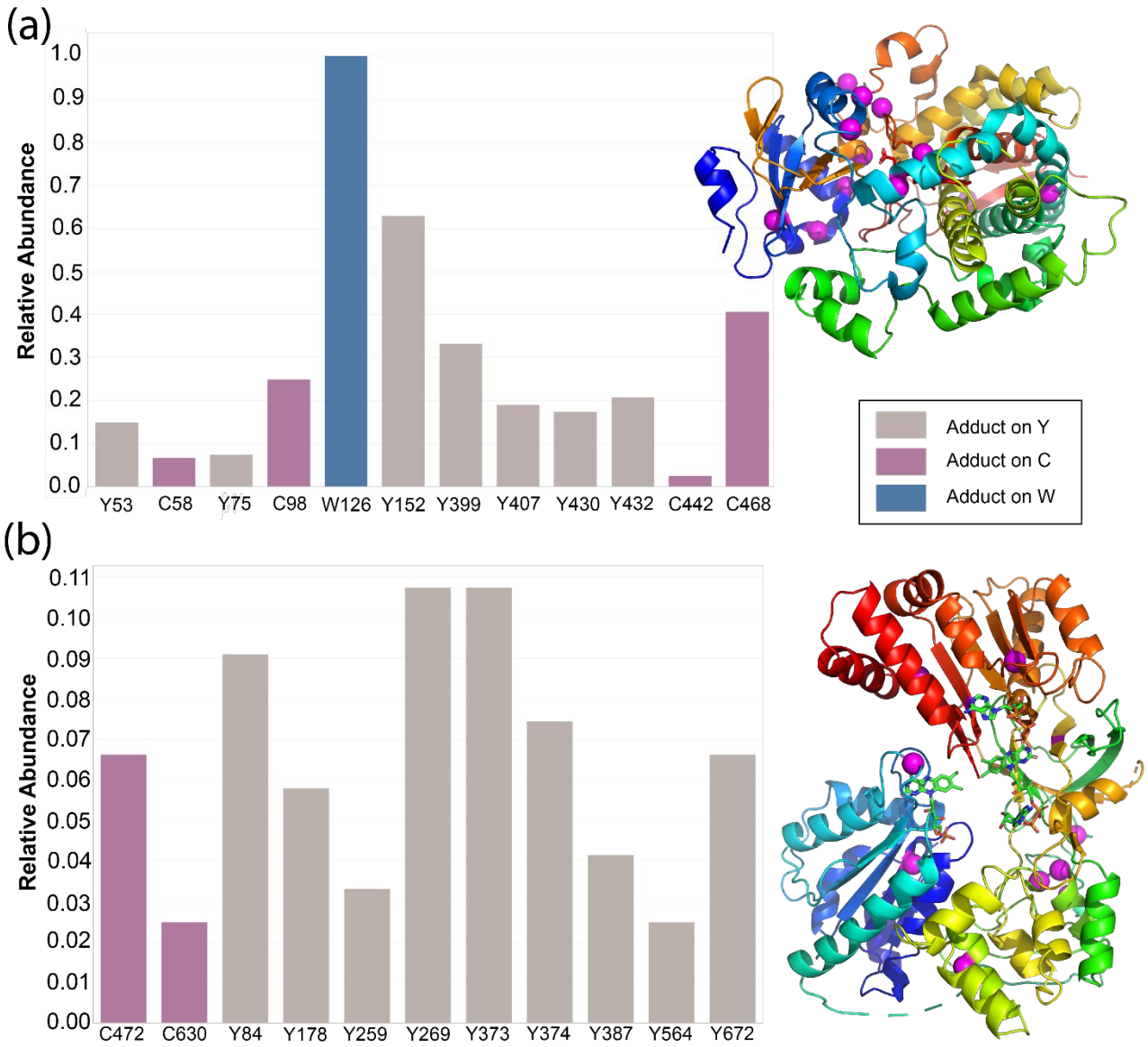


Figure S30: Identification of novel raloxifene adducts in (a) CYP3A4 and (b) P450-reductase. Abundance of all modifications is shown based on the number of PSMs identified in each location relative to CYP3A4 W126, which was observed in 121 PSMs (see Table S13 above). Observed 471 Da modifications are shown on the structures of CYP3A4 (PDB: 1TQN) and P450-reductase (PDB: 1AMO) as magenta spheres. Supplementary Files 3 contains pymol sessions of these images. All results shown have a Percolator calculated PSM level q≤0.01 and were identified by ≥3 PSMs. Raw data are available on Limelight at: https://limelight.yeastrc.org/limelight/go/0UjwIJNz45 and https://limelight.yeastrc.org/limelight/go/ypAqoCB3IE

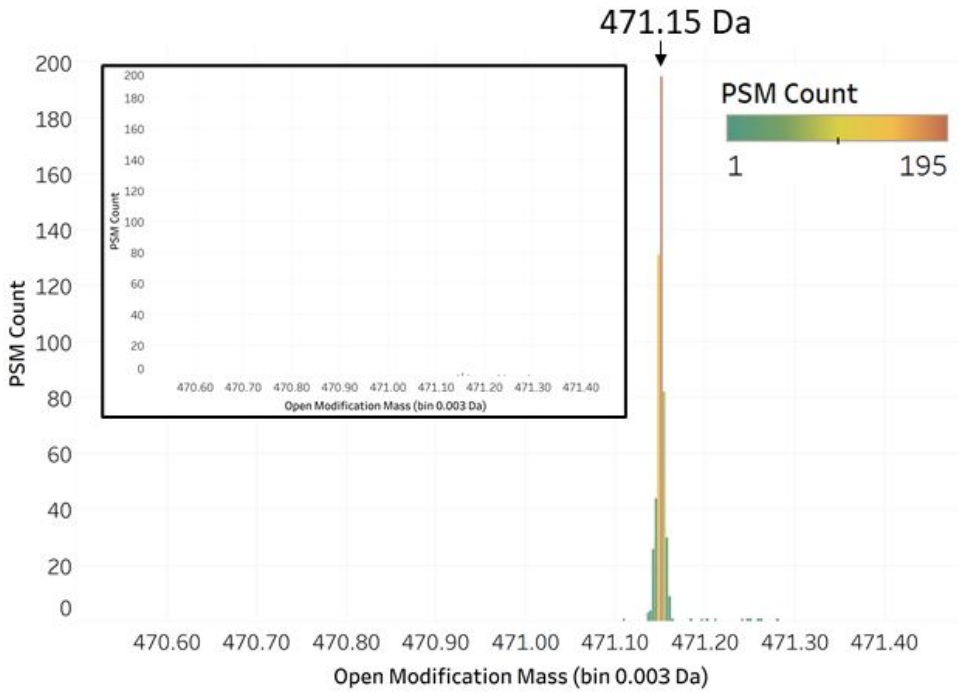


Figure S31: Distribution of Magnum identified open-mass modifications within the 471 Da bin in all untreated and raloxifene treated samples combined. Open modification masses of the 536 PSMs identified in the treated samples peak at 471.15 Da, the exact mass of the previously characterized raloxifene diquinone methide adduct. Untreated samples only resulted in 7 PSMs containing an open modification mass in the 471 Da bin at 1% FDR (raw data available can be viewed on Limelight here: https://limelight.yeastrc.org/limelight/go/jxDMcy25nY (untreated 471 Da PSMs) and here https://limelight.yeastrc.org/limelight/go/8GRABR2Ecd (treated 471 Da PSMs).

Verification of raloxifene adducts by standard closed searching

The raloxifene adducts identified in these studies were previously undiscovered and include modifications to amino acid residues (Y, W) not previously observed as modified by raloxifene. We therefore include additional verification of these identifications in the form of results from the popular closed search comet-Percolator pipeline.

All raloxifene LC-MS/MS datasets were searched using comet allowing a variable modification of mass 471.1504 Da on C, W or Y in addition to oxidation of M and alkylation of C as variable modifications. The results of these searches can be viewed on limelight here: https://limelight.yeastrc.org/limelight/go/oMwpg7cPKc and corroborate the results of our Magnum-based open searches.

A complete list of peptides found by comet to be modified by 471.1504 Da in CYP3A4 at a Percolator assigned peptide level q ≤ 0.01 can be viewed here:
https://limelight.yeastrc.org/limelight/go/bs5EzcQJEO

A complete list of peptides found by comet to be modified by 471.1504 Da in rat P450-reductase at a Percolator assigned peptide level q ≤ 0.01 can be viewed here:
https://limelight.yeastrc.org/limelight/go/nDhvoBn9As

All 471 Da modifications reported by Magnum in Table S13 were also identified by a closed comet search filtered at a 1% false discovery rate. Additional raloxifene adducts were found by this refined search method and can be viewed using the links above but are otherwise not discussed.

Representative automatically annotated spectra of raloxifene adducts identified by comet on CYP3A4 and rat P450-reductase are provided in table below. These spectra constitute comet verification of the Magnum results reported in Table S13 from our open search results.

Table S14: Comet verification of Magnum-identified 471 Da adducts presented in Table S12. All results shown have a Percolator calculated peptide level q ≤ 0.01. Links to representative spectra are provided. Full raw and data and visualizations are available on Limelight at using the links in the paragraph above.

Protein	PDB numbering modification location	Fasta numbering modification location (used in Limelight)	Annotated representative spectrum from a closed comet search defining a 471.1504 Da raloxifene adduct mass
CYP3A4	**Y53**	43	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3169/psm/196920543
CYP3A4	**C58**	48	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2344/psm/140285049
CYP3A4	**Y75**	65	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2343/psm/140201004
CYP3A4	**C98**	88	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2344/psm/140254102
CYP3A4	**W126**	116	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2344/psm/140295133
CYP3A4	**Y152**	142	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3172/psm/197086589
CYP3A4	**Y399**	389	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3170/psm/196964656
CYP3A4	**Y407**	397	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3171/psm/197032661
CYP3A4	**Y430**	420	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3172/psm/197075309
CYP3A4	**Y432**	422	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3170/psm/196982861
CYP3A4	**C442**	432	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3171/psm/197050835
CYP3A4	**C468**	458	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2343/psm/140195155
Reductase	**Y84**	84	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3169/psm/196937095
Reductase	**Y178**	178	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3171/psm/197034506
Reductase	**Y259**	259	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3172/psm/197095542
Reductase	**Y269**	269	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3172/psm/197109250
Reductase	**Y373**	373	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3169/psm/196930637
Reductase	**Y374**	374	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3170/psm/196976455
Reductase	**Y387**	387	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3169/psm/196944822
Reductase	**C472**	472	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2343/psm/140237473
Reductase	**Y564**	564	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3170/psm/196997375
Reductase	**C630**	630	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/2343/psm/140207314
Reductase	**Y672**	672	https://limelight.yeastrc.org/limelight/d/pg/spectrum-viewer/ps/3170/psm/196990948

Key protein sequences:

Table S15: Protein sequences of human serum albumin (HSA) plus the heterologously expressed proteins CYP3A4 and rat P450 reductase proteins.

>sp|P02768|ALBU_HUMAN Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2
MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKALVLIAFAQYLQQCPFEDHVKLVNEVTEFAKTCVADESAENC
DKSLHTLFGDKLCTVATLRETYGEMADCCAKQEPERNECFLQHKDDNPNLPRLVRPEVDVMCTAFHDNEETFLKKYLYEIARRHPY
FYAPELLFFAKRYKAAFTECCQAADKAACLLPKLDELRDEGKASSAKQRLKCASLQKFGERAFKAWAVARLSQRFPKAEFAEVSKL
VTDLTKVHTECCHGDLLECADDRADLAKYICENQDSISSKLKECCEKPLLEKSHCIAEVENDEMPADLPSLAADFVESKDVCKNYA
EAKDVFLGMFLYEYARRHPDYSVVLLLRLAKTYETTLEKCCAAADPHECYAKVFDEFKPLVEEPQNLIKQNCELFEQLGEYKFQNA
LLVRYTKKVPQVSTPTLVEVSRNLGKVGSKCCKHPEAKRMPCAEDYLSVVLNQLCVLHEKTPVSDRVTKCCTESLVNRRPCFSALE
VDETYVPKEFNAETFTFHADICTLSEKERQIKKQTALVELVKHKPKATKEQLKAVMDDFAAFVEKCCKADDKETCFAEEGKKLVAA
SQAALGL

>cu|hCYP3A4pCWOri|hCYP3A4pCWOri Cytochrome P450 3A4 cloned in pCWOri OS=Homo sapiens OX=9606
MALLLAVFLVLLYLYGTHSHGLFKKLGIPGPTPLPFLGNILSYHKGFCMFDMECHKKYGKVWGFYDGQQPVLAITDPDMIKTVLVK
ECYSVFTNRRPFGPVGFMKSAISIAEDEEWKRLRSLLSPTFTSGKLKEMVPIIAQYGDVLVRNLRREAETGKPVTLKDVFGAYSMD
VITSTSFGVNIDSLNNPQDPFVENTKKLLRFDFLDPFFLSITVFPFLIPILEVLNICVFPREVTNFLRKSVKRMKESRLEDTQKHR
VDFLQLMIDSQNSKETESHKALSDLELVAQSIIFIFAGYETTSSVLSFIMYELATHPDVQQKLQEEIDAVLPNKAPPTYDTVLQME
YLDMVVNETLRLFPIAMRLERVCKKDVEINGMFIPKGVVVMIPSYALHRDPKYWTEPEKFLPERFSKKNKDNIDPYIYTPFGSGPR
NCIGMRFALMNMKLALIRVLQNFSFKPCKETQIPLKLSLGGLLQPEKPVVLKVESRDGTVSGASTHHHHHH

>sp|P00388|NCPR_RAT NADPH--cytochrome P450 reductase OS=Rattus norvegicus OX=10116 GN=Por PE=1 SV=3
MGDSHEDTSATMPEAVAEEVSLFSTTDMVLFSLIVGVLTYWFIFRKKKEEIPEFSKIQTTAPPVKESSFVEKMKKTGRNIIVFYGS
QTGTAEEFANRLSKDAHRYGMRGMSADPEEYDLADLSSLPEIDKSLVVFCMATYGEGDPTDNAQDFYDWLQETDVDLTGVKFAVFG
LGNKTYEHFNAMGKYVDQRLEQLGAQRIFELGLGDDDGNLEEDFITWREQFWPAVCEFFGVEATGEESSIRQYELVVHEDMDVAKV
YTGEMGRLKSYENQKPPFDAKNPFLAAVTANRKLNQGTERHLMHLELDISDSKIRYESGDHVAVYPANDSALVNQIGEILGADLDV
IMSLNNLDEESNKKHPFPCPTTYRTALTYYLDITNPPRTNVLYELAQYASEPSEQEHLHKMASSSGEGKELYLSWVVEARRHILAI
LQDYPSLRPPIDHLCELLPRLQARYYSIASSSKVHPNSVHICAVAVEYEAKSGRVNKGVATSWLRAKEPAGENGGRALVPMFVRKS
QFRLPFKSTTPVIMVGPGTGIAPFMGFIQERAWLREQGKEVGETLLYYGCRRSDEDYLYREELARFHKDGALTQLNVAFSREQAHK
VYVQHLLKRDREHLWKLIHEGGAHIYVCGDARNMAKDVQNTFYDIVAEFGPMEHTQAVDYVKKLMTKGRYSLDVWS

Supplementary References:

(1) Woods, C. M.; Fernandez, C.; Kunze, K. L.; Atkins, W. M. Allosteric Activation of Cytochrome P450 3A4 by α-Naphthoflavone: Branch Point Regulation Revealed by Isotope Dilution Analysis. *Biochemistry* **2011**, *50* (46), 10041–10051. https://doi.org/10.1021/bi2013454.

(2) Redhair, M.; Hackett, J. C.; Pelletier, R. D.; Atkins, W. M. Dynamics and Location of the Allosteric Midazolam Site in Cytochrome P4503A4 in Lipid Nanodiscs. *Biochemistry* **2020**, *59* (6), 766–779. https://doi.org/10.1021/acs.biochem.9b01001.

(3) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat Biotechnol* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(4) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: An Open-Source MS/MS Sequence Database Search Tool. *Proteomics* **2013**, *13* (1), 22–24. https://doi.org/10.1002/pmic.201200439.

(5) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J.; Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923–925. https://doi.org/10.1038/nmeth1113.

(6) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. IProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), 1–16. https://doi.org/10.1074/mcp.M111.007690.

(7) Shteynberg, D. D.; Deutsch, E. W.; Campbell, D. S.; Hoopmann, M. R.; Kusebauch, U.; Lee, D.; Mendoza, L.; Midha, M. K.; Sun, Z.; Whetton, A. D.; Moritz, R. L. PTMProphet: Fast and Accurate Mass Modification Localization for the Trans-Proteomic Pipeline. *J. Proteome Res.* **2019**, *18* (12), 4262–4272. https://doi.org/10.1021/acs.jproteome.9b00205.

(8) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(9) Schilling, B.; Rardin, M. J.; MacLean, B. X.; Zawadzka, A. M.; Frewen, B. E.; Cusack, M. P.; Sorensen, D. J.; Bereman, M. S.; Jing, E.; Wu, C. C.; Verdin, E.; Kahn, C. R.; MacCoss, M. J.; Gibson, B. W. Platform-Independent and Label-Free Quantitation of Proteomic Data Using MS1 Extracted Ion Chromatograms in Skyline: Application to Protein Acetylation and Phosphorylation. *Mol. Cell. Proteomics* **2012**, *11* (5), 202–214. https://doi.org/10.1074/mcp.M112.017707.

(10) Keller, A.; Eng, J.; Zhang, N.; Li, X. jun; Aebersold, R. A Uniform Proteomics MS/MS

Analysis Platform Utilizing Open XML File Formats. *Mol. Syst. Biol.* **2005**, *1*. https://doi.org/10.1038/msb4100024.

(11) Hoopmann, M. R.; Zelter, A.; Johnson, R. S.; Riffle, M.; MacCoss, M. J.; Davis, T. N.; Moritz, R. L. Kojak: Efficient Analysis of Chemically Cross-Linked Protein Complexes. *J. Proteome Res.* **2015**, *14* (5), 2190−2198. https://doi.org/10.1021/pr501321h.

(12) Eng, J. K.; Hoopmann, M. R.; Jahan, T. A.; Egertson, J. D.; Noble, W. S.; MacCoss, M. J. A Deeper Look into Comet—Implementation and Features. *J. Am. Soc. Mass Spectrom.* **2015**, 1865–1874. https://doi.org/10.1007/s13361-015-1179-x.

(13) Eng, J. K.; McCormack, A. L.; Yates 3rd, J. R. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J Am Soc Mass Spectrom* **1994**, *5*, 976–989.

(14) Eng, J. K.; Hoopmann, M. R.; Jahan, T. A.; Egertson, J. D.; Noble, W. S.; Maccoss, M. J. A Deeper Look into Comet — Implementation and Features. **2015**, 1865–1874. https://doi.org/10.1007/s13361-015-1179-x.

(15) Eng, J. K.; Fischer, B.; Grossmann, J.; MacCoss, M. J. A Fast SEQUEST Cross Correlation Algorithm. *J. Proteome Res.* **2008**, *7* (10), 4598–4602. https://doi.org/10.1021/PR800420S.

(16) Noble, W. S. Mass Spectrometrists Should Search Only for Peptides They Care About. *Nat. Methods* **2015**, *12* (7), 605–608. https://doi.org/10.1038/nmeth.3450.

(17) Sharma, V.; Eng, J. K.; Maccoss, M. J.; Riffle, M. A Mass Spectrometry Proteomics Data Management Platform. *Mol Cell Proteomics* **2012**, *11* (9), 824–831. https://doi.org/10.1074/mcp.O111.015149.

(18) Yi, X.; Gong, F.; Fu, Y. Transfer Posterior Error Probability Estimation for Peptide Identification. *BMC Bioinformatics* **2020**, *21* (1), 1–17. https://doi.org/10.1186/s12859-020-3485-y.

(19) Vizcaíno, J. A.; Côté, R.; Reisinger, F.; Foster, J. M.; Mueller, M.; Rameseder, J.; Hermjakob, H.; Martens, L. A Guide to the Proteomics Identifications Database Proteomics Data Repository. *Proteomics*. Wiley-Blackwell September 2009, pp 4276–4283. https://doi.org/10.1002/pmic.200900402.

(20) Parker, C. E.; Perkins, J. R.; Tomer, K. B. Nanoscale Packed Capillary Liquid Chromatography- Electrospray Ionization Mass Spectrometry : Analysis of Penicillins and Cephems. **1993**, *616*, 45–51.

(21) Jenkins, R. E.; Meng, X.; Elliott, V. L.; Kitteringham, N. R.; Pirmohamed, M.; Park, B. K. Characterisation of Flucloxacillin and 5-Hydroxymethyl Flucloxacillin Haptenated HSA in Vitro and in Vivo. *Proteomics - Clin. Appl.* **2009**, *3* (6), 720–729. https://doi.org/10.1002/prca.200800222.

(22) Creasy, D. M.; Cottrell, J. S. Unimod: Protein Modifications for Mass Spectrometry. *Proteomics*. John Wiley & Sons, Ltd June 1, 2004, pp 1534–1536. https://doi.org/10.1002/pmic.200300744.

(23) Lawrence, R. T.; Searle, B. C.; Llovet, A.; Villén, J. Plug-and-Play Analysis of the Human Phosphoproteome by Targeted High-Resolution Mass Spectrometry. *Nat. Methods* **2016**, *13* (5), 431–434. https://doi.org/10.1038/nmeth.3811.

(24) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining Results of Multiple Search Engines in Proteomics. *Mol. Cell. Proteomics* **2013**, *12* (9), 2383–2393. https://doi.org/10.1074/mcp.R113.027797.

(25) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; Eng, J. K.; Martin, D. B.; Nesvizhskii, A. I.; Aebersold, R. A Guided Tour of the Trans-Proteomic Pipeline. *Proteomics*. NIH Public Access March 2010, pp 1150–1159. https://doi.org/10.1002/pmic.200900375.

(26) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; MacCoss, M. J.; Noble, W. S. Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. *J. Proteome Res.* **2014**, *13* (10), 4488–4491. https://doi.org/10.1021/pr500741y.

(27) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(28) Chi, H.; Liu, C.; Yang, H.; Zeng, W. F.; Wu, L.; Zhou, W. J.; Wang, R. M.; Niu, X. N.; Ding, Y. H.; Zhang, Y.; Wang, Z. W.; Chen, Z. L.; Sun, R. X.; Liu, T.; Tan, G. M.; Dong, M. Q.; Xu, P.; Zhang, P. H.; He, S. M. Comprehensive Identification of Peptides in Tandem Mass Spectra Using an Efficient Open Search Engine. *Nat. Biotechnol.* **2018**, *36* (11), 1059–1066. https://doi.org/10.1038/nbt.4236.

(29) Bagwan, N.; Bonzon-kulichenko, E.; Calvo, E.; Lechuga-vieco, A. V.; Michalakopoulos, S.; Trevisan-Herraz, M.; Ezkurdia, I.; Rodrıiguez, J.; Magni, R.; Latorre-pellicer, A.; Enrı, A.; Enriquez, J.; Vazquez, J. Comprehensive Quantification of the Modified Proteome Reveals Oxidative Heart Damage in Mitochondrial Heteroplasmy. *Cell Rep.* **2018**, *23*, 3685–3697. https://doi.org/10.1016/j.celrep.2018.05.080.

(30) Solntsev, S. K.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Enhanced Global Post-Translational Modification Discovery with MetaMorpheus. *J. Proteome Res.* **2018**, *17*, 1844–1851. https://doi.org/10.1021/acs.jproteome.7b00873.

(31) Na, S.; Bandeira, N.; Paek, E. Fast Multi-Blind Modification Search through Tandem Mass Spectrometry. *Mol. Cell. Proteomics* **2012**, *11* (4), M111.010199. https://doi.org/10.1074/mcp.M111.010199.

(32) Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E. TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Datasets. *Nat. Biotechnol.* **2019**, *37* (4), 469–479. https://doi.org/10.1038/s41587-019-0067-5.

(33) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.

(34) Li, Q.; Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Schaffer, L. V.; Scalf, M.; Smith, L. M. Global Post-Translational Modification Discovery. *J. Proteome Res.* **2017**, *16* (4), 1383–1390. https://doi.org/10.1021/acs.jproteome.6b00034.

(35) David, M.; Fertin, G.; Rogniaux, H.; Tessier, D. SpecOMS: A Full Open Modification Search Method Performing All-to-All Spectra Comparisons within Minutes. *J. Proteome*

Res. **2017**, *16*, 3030–3038. https://doi.org/10.1021/acs.jproteome.7b00308.

(36) Pino, L. K.; Searle, B. C.; Bollinger, J. G.; Nunn, B.; Maclean, B.; Maccoss, M. J. The Skyline Ecosystem: Informatics for Quantitative Mass Spectrometry Proteomics. *Mass Spectrom. Rev.* **2017**. https://doi.org/10.1002/mas.21540.

(37) Baer, B. R.; Wienkers, L. C.; Rock, D. A. Time-Dependent Inactivation of P450 3A4 by Raloxifene: Identification of Cys239 as the Site of Apoprotein Alkylation. *Chem. Res. Toxicol.* **2007**, *20* (6), 954–964. https://doi.org/10.1021/tx700037e.